PhD degree in Foundations of the Life Sciences and their Ethical Consequences

European School of Molecular Medicine (SEMM) and University of Milan
Faculty of Medicine

Settore disciplinare: FIL/02

Humans, animals, and Petri dishes: Biomedical modeling between experimentation and representation

Pierre-Luc Germain

IFOM-IEO Campus, Milan
Matricola n. R08889

Internal Supervisor & Lab Supervisor:	Giuseppe Testa, European Institute of Oncology IFOM-IEO Campus, Milan, Italy
External Supervisor:	Marcel Weber, University of Geneva Geneva, Switzerland

Anno accademico 2012-2013

Contents

Introduction **1**
0.1 Biomedical models as a working category . . . 2
0.2 Descriptivity and normativity . . . 4
0.3 The context of enquiry . . . 6
0.3.1 Ethics of animal experimentation . . . 6
0.3.2 Funding policy and research prioritization . . . 8
0.4 Plan . . . 9

1 Do biomedical models model? **13**
1.1 Philosophical accounts of scientific models . . . 14
1.1.1 Models as mediators . . . 15
1.1.2 Models as representations . . . 16
1.1.3 Similarity and adequacy of representation . . . 18
1.1.4 Modeling as surrogacy or indirect access . . . 20
1.1.5 Model of, model for . . . 24
1.2 Surrogacy in biomedical models . . . 27
1.2.1 Cutting lineages at the joints . . . 28
1.2.2 Samples in mosaic individuals . . . 30
1.2.3 Material idealizations . . . 33
1.2.4 Abandoning the model/sample distinction in biomedical research . 35
1.3 Animal models, model animals . . . 36
1.3.1 Model individuation . . . 38
1.4 Models or experimental systems? . . . 41

2 Evaluating biomedical models **45**
2.1 Introduction . . . 45
2.2 Screens and predictivity . . . 47
2.2.1 The critique of animal testing . . . 47
2.2.2 Multi-step screening . . . 50
2.2.3 The missing base rate . . . 53
2.2.4 Sensitivity and specificity . . . 55
2.3 The Cancer Chemotherapy National Service Center . . . 57
2.3.1 Screening in the context of drug discovery . . . 57
2.3.2 The CCNSC – from attrition to annotation . . . 60
2.4 Assumptions of the received views . . . 66
2.4.1 Assumption 1: Biomedical models function as surrogates . . . 66
2.4.2 Assumption 2: Modeling is a unidirectional process . . . 69
2.4.3 Assumption 3: Modeling is a dyadic relationship . . . 70
2.4.4 A new account of biomedical models . . . 71

3 The instrumental role of biomedical models 73
3.1 Introduction . . . 73
3.1.1 Kinds of kinds of biomedical models . . . 74
3.2 Living instruments . . . 76
3.2.1 Example 1: The Ascheim-Zondek test for pregnancy . . . 78
3.2.2 Example 2: Induced pluripotent stem cell models of cancer . . . 79
3.2.3 Example 3: Xenograft models of cancer . . . 81
3.2.4 Example 4: Genetically engineered models of cancer . . . 85
3.3 Characterization of the instrumental role . . . 89
3.3.1 Production and observation instruments . . . 91
3.3.2 Replica and instruments . . . 95
3.3.3 The growing importance of the instrumental role . . . 98
3.4 Proximate functions of biomedical models . . . 100
3.4.1 Models as tools for thinking . . . 100
3.4.2 Models as replica or surrogates . . . 101
3.4.3 Models as instruments . . . 101
3.4.4 Models as reagents or factories . . . 101
3.5 Conclusion . . . 102

4 Models and theory 105
4.1 Introduction . . . 105
4.2 Extrapolation and the theoretical . . . 107
4.2.1 Claude Bernard and the foundations of experimental medicine . . . 107
4.2.2 Theoretical constructs . . . 111
4.2.3 Theoretical constructs as bridges . . . 115
4.3 Cancer stem cells between theory and operations . . . 116
4.3.1 The theoretical grounding of early cancer xenografts . . . 116
4.3.2 Epistemic iteration . . . 120
4.3.3 The Cancer Stem Cell framework . . . 121
4.3.4 Melanoma-initiating cells . . . 124
4.4 Spaces of representation . . . 128
4.4.1 The closure of representation . . . 130
4.5 Conclusion . . . 132

5 In vitro models and distributed modeling 135
5.1 Introduction . . . 135
5.2 Stem cell models . . . 136
5.2.1 A brief history of iPSCs . . . 136
5.2.2 Repeated development . . . 138
5.2.3 Cell types between in vivo and in vitro . . . 143
5.2.4 Turning the inside out and recomposing it . . . 146
5.3 The iPSC research paradigm . . . 149
5.3.1 iPSC models and human variation . . . 149
5.3.2 In vitro symptoms as intermediary anchorpoints . . . 152
5.3.3 Translating phenotypes . . . 154
5.3.4 Comparative cellular models . . . 158
5.4 Distributed modeling . . . 160
5.4.1 Against the dyadic view . . . 160
5.4.2 The role of the patient . . . 163
5.4.3 Structures of models as a locus of evaluation . . . 165

5.4.4 Connectivity 166
5.5 The new model organism? 169

Conclusion **173**
6.6 Modeling in applied research 176

Appendix **179**

A Test statistics and the coin-tossing argument **179**
A.1 Fundamentals of test evaluation 179
A.2 Base rates and the coin-tossing argument 180

B Biological and technical replicates **183**

Acknowledgments **187**

Bibliography **189**

List of Abbreviations

ADR	Abstract Direct Representation
BEAR	Biological Effects of Atomic Radiation
CAM	Causal Analogical Models
CCNSC	Cancer Chemotherapy National Service Center
CERN	Conseil Européen pour la Recherche Nucléaire (European Organization for Nuclear Research)
CSC	Cancer Stem Cell
DNA	Deoxyribonucleic Acid
DTP	Developmental Therapeutics Program (of the NCI)
ESC	Embryonic Stem Cell
FACS	Fluorescence-Activated Cell Sorting
FN	False Negative
FP	False Positive
GWAS	Genome-Wide Association Study
HAM	Hypothetical Analogical Model
HGNC	Human Genome Organization Gene Nomenclature Consortium
hESC	Human Embryonic Stem Cell
iPSC	Induced Pluripotent Stem Cell
LPS	Lipopolysaccharide
MGI	Mouse Genome Institute
MIC	Melanoma-Initiating Cell
mRNA	Messenger RNA (Ribonucleic Acid)
NCI	National Cancer Institute (of the United States of America)
NEP	Neuroepithelial Progenitors
NOD/SCID	Non-Obese Diabetic/Severe Combined Immuno-Deficiency
NPV	Negative Predictive Value
NRC	National Research Council (of the United States of America)
NSC	Neural Stem Cell
PBPK	Physiologically-Based Pharmacokinetic
PPV	Positive Predictive Value
PSI-MI	Proteomics Standards Initiative for Molecular Interactions
RDoc	Research Domain Criteria
RNA	Ribonucleic Acid
SVZ	Subventricular Zone

TALEN Transcription Activator-Like Effector Nucleases
TN True Negative
TP True Positive
WBS Williams-Beuren Syndrome
7Dup 7q11 micro-Duplication Syndrome

Gene names / products

ABCB5 ATP (adenosine triphosphate)-Binding Cassette, sub-family B, member 5
cMyc Cellular Myelocytomatosis oncogene (HGNC symbol: MYC)
FMR1 Fragile-X Mental Retardation 1
HER2 Human Epidermal Growth Factor Receptor 2 (HGNC symbol: ERBB2)
HRAS Harvey Rat Sarcoma Viral oncogene homolog
Il2rg Interleukin-2 Gamma Receptor
LRRK2 Leucine-Rich Repeat Kinase 2
SOX2 SRY (sex determining region Y)-box 2

List of Figures

2.1 The CCNSC pipeline 63

3.1 The larval phenotype of the zebrafish model 86

5.1 *In vitro* cerebral organoid 141

5.2 Correspondence between *in vivo* and *in vitro* 144

5.3 The iPSC research platform 153

5.4 Translation of disease phenotypes 157

Abstract

The use and evaluation of biomedical models – *in vitro* and *in vivo* models of diseases – is discussed among scientists, philosophers and science policy-makers, in an attempt to maximize the efficiency and relevance of research and minimize unnecessary moral costs. However, such evaluations raise several methodological issues and have generally been hampered by a lack of attention to the precise functions played by biomedical models. For if biomedical models are ultimately expected to inform us about human pathologies, they seldom do so in isolation, and get there through a wide variety of ways. An epistemological understanding of this process is therefore a precondition for their evaluation, and this thesis is an attempt at building such an epistemology.

Several of the examples used come from cancer research, especially xenograft models and models used in the context of large-scale drug screenings. Another important set of examples come *in vitro* models, with a particular focus on disease modeling using induced pluripotent stem cells.

I argue that the notion of model, if conceived as to apply to biomedical models, conflates into that of experimental system. I therefore propose an account of biomedical models that does not presuppose a fundamental divide between modeling and experimentation. I show that biomedical models are not simply scaled-down versions of their target, but instead projections of their target in a different space of representation. I argue for an instrumental role of biomedical models, and use this role to explore the diversity of proximal functions fulfilled by biomedical models. I propose the notion of distributed modeling to draw attention to the relations between model systems, and illustrate this by analyzing the interplay between *in vitro* and *in vivo* models. Finally, I explore the implications of this account for the evaluation of biomedical models, and more broadly for the topic of scientific representation.

x

Introduction

Biomedical research ultimately aims at improving our understanding of human biology and pathology, and our capacity to intervene on it. However, a considerable proportion of this research proceeds in the absence of its main object – in the absence of human patients. It is the main goal of this thesis to understand how this is possible, and to provide an epistemology for biomedical models.

I want to draw an account of biomedical models that does not take for granted some important categories in the way we have come to think about biomedical research. What I call the 'directional and dyadic view of modeling' understands modeling as a dyadic relationship between a model and a target system, in which facts are first worked out in the model before being extrapolated to the target. This view also presupposes that modeling is fundamentally distinct from something else that remains uncharacterized – what we might call 'direct experimentation' – and tends to assume that biomedical models function as surrogates for human patients. I believe that such a view is flawed in many respects, and my goal is to highlight these shortcomings and propose an alternative view of biomedical research. In doing so, I wish not to take for granted what models and targets are, nor presuppose a division between modeling and experimentation.

I want to argue that translational research (or translational medicine), often understood as the translation of lab research into clinical application, is not different in nature from translations made between experimental systems. However, one should take seriously the metaphor of translation. As anyone who has been involved in (literal) translation knows all too well, words do not have a one-to-one equivalent in other languages. Translation requires attention to the context in which the word is used, because the meaning of a word

When quoting documents in another language, I have provided a free translation in the body of the text for ease of reading, and provided the original passage in footnote.

depends on its neighbors.

In a similar way, I will argue that biomedical models function in a variety of ways but very seldom on their own. For this reason, I propose an account of biomedical research that pays close attention to the relationship model systems (including clinical systems) entertain with each other. In saying that biomedical models work together, I do not simply mean that they are complementary, but that they are synergistic. To understand this, I propose the notion of 'distributed modeling', and explore its consequences for translational research and the evaluation of biomedical models.

0.1 Biomedical models as a working category

Most researchers do not conduct studies on human beings, and instead work on so-called 'models'. Animals have been used in this way at least since antiquity, and more recently *in vitro* and *in silico* models have been added to the toolkit of biomedical research. *In silico* models stands for computer models and simulations. The notion of *in vitro* is instead more complex, and its meaning changes across historical periods and scientific field, and indeed even withing the daily practice of a single scientist. Despite its etymology, it is not defined by the test tube but negatively, sometimes meaning outside (in the absence) of whole cells, and sometimes meaning outside of whole organisms (I will discuss this depth in Chapter 5, section 5.2.4). In biomedical research, the latter use predominates (for reasons that will become obvious), and cell culture is generally referred to as *in vitro*[1]. Unless specified otherwise, I will therefore use the expression *in vitro* in this sense, and consequently use *in vivo* in the context of whole (multicellular) organisms, most often animals.

These different kinds of 'models' – *in vitro, in vivo, in silico* models – are very heterogeneous in nature and in the practice they enable. To study them all would not allow the necessary depth, and for this reason this thesis will be mainly focused on *in vitro* and *in vivo* models. I will leave to future work the task of including *in silico* models in this account.

There are a number of reasons for this choice. First, among the three, *in vitro* and *in vivo* models can be grouped together in that they are both material (concrete), biological

[1] With the exception of primary cell culture (cells recently derived from an organism), which are commonly called *ex vivo*. The distinction between *ex vivo* and *in vitro* is obviously very blurry.

(living) models which to some extent embody the phenomena of interest. Instead, *in silico* models are arguably more akin to traditional abstract models. Moreover, *in vitro* and *in vivo* models are more peculiar and historically more fundamental to the life sciences, and are therefore more likely to reveal some of its specificities, than *in silico* models. The latter are instead a very recent addition to most, if not all fields of scientific research. Finally, it is one of the claims of this thesis that *in vitro* and *in vivo* models have become meshed in inextricable ways, and therefore ought to be studied together (see Chapter 5). While the same can be said locally of *some in silico* models, I believe that more fundamental insights on biomedical research in general is still to be gained even without systematic attention to its *in silico* components.

Throughout the present work, and unless specified otherwise, I therefore will use the expression 'biomedical models' for *in vitro* and *in vivo* systems which are used in the course of biomedical research. This definition will appear to most philosophers as much too broad, most prominently because it seems not to involve any modeling relationship. An important reason for this choice is that understanding the concept of model is not, for me, an end in itself, but merely a potential means of understanding biomedical research. To this end, it is counter-productive to presuppose a strict philosophical account of models which, as I show in Chapter 1, will ultimately turn out to be very problematic when applied to concrete biological systems. I will instead suggest that in the context of the concrete biological systems used in biomedical research, epistemological analysis is best served by more refined categories sensitive to the precise role played by these systems.

Finally, it should be immediately clear that this is not a thesis about 'model organisms', and that it touches the topic only indirectly. Biomedical models include animal models, some of which will involve model organisms. However, philosophical discussion on model organisms have generally been framed around their representativity for a broader range of species, and the value of (or issues related to) the concentration of biological research around model organisms (Bolker 1995, Burian 1993, Leonelli 2007, Ankeny and Leonelli 2011). In contrast, biomedical models are by definition meant to accomplish a different job (and I would argue that it is not just a narrower job, but indeed a *different* one). For this reason, I will approach the topic of model organisms only when and to the extent that

they are necessary to my primary concern. For a deeper exploration of the issue of model organisms in the context of biomedical research, see the works collected in Gachelin (2006).

0.2 Descriptivity and normativity

> "An adequate philosophy of science should have normative force." (Wimsatt 2007, p.26)

Work in philosophy of science is traditionally divided into normative and descriptive approaches. Normative philosophers have tried to defend accounts of how science ought to be done, focusing on justification and rational inference. The Vienna Circle remains today one of the most powerful illustrations of this endeavor, and yet is often perceived as remote from actual scientific practice, especially in fields such as biology. This gap has led to the widespread idea that scientists know about their trade better than philosophers in their armchair, and that philosophy should set aside its precepts for a moment and learn what science can teach about rationality and inference. This turn brought attention to then under-appreciated aspects of science, especially its practice.

Descriptive philosophy of science however received its own criticisms, summarized in the famous taunt attributed to Richard Feynman, according to which 'philosophy of science is about as useful to science as ornithology is to birds'. Although the taunt was not addressed at any particular strand of philosophy of science, the question it asks is perhaps even more urgent to the descriptive philosopher. A scientist regularly attending philosophy conferences once confided to me his distress at coming there to think critically about how to do science, only to find the philosophers in turn looking at what he was doing.

The present work is deliberately neither, or rather both, descriptive and normative. Indeed I believe that in philosophy of science, the question of whether an account is normative or descriptive is ill-posed. Philosophy of science necessarily implies both descriptivity and normativity. It implies descriptivity because it does not begin in a void, as Descartes alleged to do, but with terms, problems and concerns that are already those of science. It implies normativity because, simply put, no description is neutral (every depiction is also predication). Every commentator presents scientific episodes or issues in a certain way, emphasizing some form of coherence or another, be it theoretical, social or otherwise.

Perhaps a good analogy for these points is that of a dictionary. For those who make it, the dictionary is a descriptive endeavor: as the meaning of a word is in its usage, dictionaries generally aim at tracking usage. But they are not exhaustive, and not all usages of a word are equal. Some users, or some instances of a word, carry more weight than others. A word used in a new way by Shakespeare is different from a word used in a new (i.e. 'erroneous') way by a peasant. Dictionaries carry with them the norms of the society that has made them, and this is even more obvious for languages which have an active regulatory institution. Furthermore, once made the dictionary is received as normative, for we look to it to know how we should spell words, and how we should use them.

Conceptual explication is an important task for philosophy of science, and it can have the same normative consequences other such efforts have. For instance, the mathematical description of a known and qualitatively simple relationship might seem of little use for it 'merely' describes the relationship in a different language, but it can also teach us where this simplicity might break down. In a similar way, philosophical description – with its abstraction, but also with its aim of making explicit the tacit and the implicit of scientific practice and discourse – can call attention to the limits of the concepts, tools and strategies science regularly uses. In analogy to William Wimsatt's self-description as a "conceptual engineer, not a pure theoretician" (Wimsatt 2007, p.30), the task I have appointed myself here is an attempt at reverse-engineering biomedical research. Reverse-engineering is not just description, but identification of what makes something function, how and why. Like much of biology, it is a work of purification. And much like in biology, what is to be purified itself changes in the course of the process.

Finally, the present enterprise has another layer of normativity, tied to any functional analysis. An account of how research uses and relies on biomedical models, because it defines their function, necessarily implies the norms according to which they are evaluated. Indeed, perhaps the main motivation for this thesis is to understand and highlight what must be taken into account for an evaluation of biomedical models.

0.3 The context of enquiry

0.3.1 Ethics of animal experimentation

Ever since the National Academy of Sciences was asked, in 1896, to "express an opinion as to the scientific value of experiments upon the lower animals and as to the probable effect of restrictive legislation upon the advancement of biological science", the debate regarding animal experiments was largely framed as a trade-off between a "trifling amount of animal suffering" and "incalculable benefits to the human race" (the letter is reprinted in Committee on the Use of Laboratory Animals 1988, p.89). How trifling the suffering is, and how important the benefits are, has been the subject of most debates regarding the use of animals. It must be noted, however, that this is not the only way to frame the issue. Indeed, it is very much *unlike* the debates on human experimentation, where individual dignity was traditionally placed above any cost-benefit consideration. The same argumentative pattern sometimes arises in public discussions. For instance, during the writing of this work, Italy's scientific community has been shaken by the Italian Parliament's approval of amendments (see Disegno di legge 587, Art. 13) to the European directive on the protection of animals used in science (2010/63/UE). Among other things, the law would completely prohibit xenotransplantation, which threatens to cripple cancer research all across the country. Beyond (unsupported) claims that animal experiments are unnecessary, proponents of the law have argued that the issue is not a scientific one (nor even a political one for that matter), but a moral one. While there is no doubt that the issue is not purely scientific, for no science will ever tell us the relative value of lives, it is far from straightforward to show that the issue is not *at all* scientific. Even Peter Singer, probably the most famous philosopher defending animal rights, frames the issue as a trade-off between interests (Singer 1975), which necessitates an evaluation of the benefits brought by animal experimentation. Declaring the issue purely non-scientific implies to abandon consequentialism[2]. Leaving that issue to moral philosophers, the problem I am interested

[2] Furthermore, an in principle ban on animal experimentation on the ground of animal rights would have a hard time justifying its application specifically to science, and would be incoherent with the considerable amounts of animals used for non-scientific purposes. This was for instance noted in the executive summary of the 1988 report of the National Research Council on the use of laboratory animals: "The quantities are a small fraction of the total of over 5 billion animals used annually for food, clothing, and other purposes in the United States." (Committee on the Use of Laboratory Animals 1988, p.2)

in is the following: assuming that the ethical question is, in fact, one of trade-off (which therefore implies both scientific and normative components), how are we to evaluate the epistemic value of animal experimentation?

In 1896, the answer of the National Academy of Sciences was straightforward, requiring barely a couple of pages:

> "That animals must suffer and die for the benefit of mankind is a law of nature [...] If this work is interfered with, medical science will continue to advance [...] but there will be this important difference, that the experimenters will be medical practitioners and the victims human beings." (Committee on the Use of Laboratory Animals 1988, p.91)

Such a call to natural order would be unconvincing in the contemporary scientific community, but its logic has continued to tip the balance throughout the 20th century. There were some waves of public concern, but the most important reconsideration of the issue happened in the second half of the century, following the burgeoning techniques of tissue culture[3]. From that moment on, the trade-off became more than a dilemma between human and animal victims, for there seemed to be another alternative: cells and tissues could replace animals in the lab. This moved the discussion to a comparison of the effectiveness of *in vitro* and *in vivo* research, and the question of whether any difference offsets the costs in animal suffering.

Contemporary scholars criticising animal experimentation (LaFollette and Shanks 1996, Shanks and Greek 2009, Knight 2011) follow this general argumentative pattern. These critics have argued against the efficacy of animal testing both on theoretical grounds and relying on empirical evidence supporting (see especially Knight 2011). The critique is however flawed in two related respects. First, as I will show in Chapter 2, it relies on a picture of animal experimentation which fits little of contemporary science. Second, these critics generally fail to conduct the further steps of the evaluation: namely to compare the poor efficiency of animal studies with its alternatives. The more fundamental problem is

[3] The social pressure on the scientific community is extremely visible in this preface (by George T. Harrell) to the proceedings of a 1975 symposium organized by the Institute of Laboratory Animal Resources (ILAR), of the US National Research Council: "The impetus for holding the symposium came because of general public interest in all facets of health, the attention given in the press to use of animals in research, the concern of Congress with the progress of research it has funded over the years, and the hope of groups of citizens concerned with animal welfare that sufficient progress has been made that changes in research project designs might be instituted in the extent and type of use of animals. The latter group spoke first to express their concern. Scientists followed with detailed discussions of the current state of the art in their own disciplines." (Institute of Laboratory Animal Resources (ILAR) 1977, p.IV)

a failure to properly consider the precise functions of the models they evaluate. Instead, I believe that any evaluation first requires a careful understanding of these functions, and propose an approach this study.

0.3.2 Funding policy and research prioritization

Although LaFollette and Shanks (1996) made their epistemic argument against animal experimentation in the context of the ethical debate on animal rights, Shanks and Greek (2009) explicitly detached themselves from these issues and considered only the question of whether our interests are well served by animal experimentation. Indeed, the ethical debates surrounding animal experimentation represent only a small portion of the discussion on the value of different biomedical models. As science took an industrial scale throughout the 20th century, starting with agriculture and spreading to biology and medicine especially after the Second World War, the choice of experimental systems stopped being left to the fancy of scientists. Biological research was gradually concentrated on a few model organisms[4]. Even funding agencies which did not have a top-down planning of scientific research needed a rationale to discriminate between research projects. This necessarily prompted discussions about the relative appropriateness of different research organisms or experimental systems. This is well illustrated by the cancer drug screening programme of the National Cancer Institute, discussed in Chapter 2 (section 2.3). The problem has seemed even more urgent in the last decades, which have seen a dramatic decline in the efficiency of drug discovery: "the number of new drugs approved per billion US dollars spent on R&D [Research & Development] has halved roughly every 9 years since 1950, falling around 80-fold in inflation-adjusted terms." (Scannell et al. 2012, p.191) As these authors note, there can be a number of reasons for this. To start with, in order to be approved a drug must in general be shown to be better than the standard of care, which means that the search criteria are progressively more stringent with time. This is exacerbated by society's lower tolerance to risk, tightening (perhaps rightly) the requirements for clinical testing and approval of drugs. In any case, biomedical research is in an efficiency crisis: it has failed

[4] See Churchill (1997) or Logan (2002) for historical case studies and discussions on decreases in laboratory organisms diversity. Bruno Strasser has argued that the last few decades have seen a return of organismal diversity in the life sciences. For more general philosophical discussion on model organisms, see especially Ankeny and Leonelli (2011).

to deliver what was expected from the magnitude of the investments. In this context, it is only natural to try to assess the value of the models on which it is based.

Notwithstanding the importance of the debates on animal rights, my contribution here will remain agnostic with regards to animal ethics. I am instead interested in the epistemological question of how a biomedical model can be said to be good or bad, better or worse than another. While I do not aim to do the proper evaluative work, I wish to lay down a necessary groundwork for such an endeavor by proposing a more adequate epistemology of disease modeling. I want to argue that the lack of attention to the functions of biomedical models has lead to a bias in their evaluation, most obvious in the obsession for the phenocopy of clinical phenomena.

0.4 Plan

Chapter 1 discusses some of the most important philosophical notions of model and their application to biomedical models. Although many brief examples are used, it is a rather theoretical chapter in which I make some general observations that will be elaborated throughout the following chapters. A first major theme is that the difference between models and samples, or between modeling and what one might call 'direct experimentation', is very blurry. I argue that the notion of model, if construed as to apply to biomedical models, collapses into that of experimental system, and as a consequence I suggest to consider modeling and direct experimentation as continuous. A second thread is that even in cases where the distinction between models and non-models is clear, it is often unclear where the model stops – what is inside it, and what is outside it. Finally, I follow a number of philosophers in considering the relation of representation as dependent on a representational system, the consequences of which will be explored in the following chapters.

Chapter 2 begins by considering the critique, by a more or less consistent set of authors, of the use of animal models in biomedical research. While I grant that the predictivity of animal models is much worse than is generally held, I argue that these authors' arguments have a very restricted scope because they presuppose a narrow understanding of biomedical models which misrepresents actual biomedical research. Relying on the example of the Cancer Chemotherapy National Service Center, I show that even paradigmatic forms of

applied research such as drug screening do not fit this view. I discuss model evaluation in relationship to its usage and to its alternatives, especially in the context of multi-step screening. Finally, I describe a set of problematic assumptions which have hampered most accounts of biomedical models: that biomedical models are surrogates for human beings, that modeling is unidirectional, and that it is a dyadic relationship between the model and the target.

Chapter 3 focuses on the first assumption identified in the previous chapter: that biomedical models function as surrogates for human patients. I present several examples of a role very different from that of a surrogate – the instrumental role, or model as measuring device. I characterize this role with respect to that of a surrogate, and argue for its relevance in contemporary biomedical research. The rationale for discussing this role is two-fold. First, because the instrumental role is so different from the surrogate role, it suggests a new and (I argue) more fruitful way to look at models, namely as projections on a different space of representation. Second, exploration of the instrumental role invites attention to the diversity of functions biomedical models accomplish. Distinguishing ultimate and proximate functions of biomedical models, I sketch a taxonomy of the latter.

Biomedical models functioning as measuring devices locate their target in a theoretically constructed space, which requires a consideration of the role of theoretical terms and frameworks in biomedical modeling. This is the topic of Chapter 4, which explores the relationship between biomedical models, extrapolation, and the theoretical realm. I consider some of the most important historical examples on the topic, starting with Claude Bernard's philosophy of medicine. I show how his justification of the validity of animal experiments was a consequence of his foundation of medicine as a science. I then turn to the notion of construct validity in psychology. In both cases, the role of theoretical constructs is to bridge material systems. I argue that this must be taken into account for the evaluation of biomedical models, especially when they are used as instruments, and illustrate this with an example from the field of Cancer Stem Cells. I relate this account to the earlier discussion of representation, and to a coherentist view on science.

Chapter 5 brings the different threads together by proposing to view biomedical research as distributed modeling. A central theme discussed in the chapter is the nature of the *in*

vitro and its relation with the *in vivo*. My main example is the use of induced pluripotent stem cells for disease modeling, of which I describe the major developments and some of the most important epistemological problems. I discuss the construction of phenomena – symptoms, or phenotypes – at the intersection between models, and the organization of multiple biomedical models (including clinical models) into a research system. I explore some implications of this account for the issue of comparing and evaluating biomedical models and propose larger modeling structures as loci of evaluation. I conclude with some more general points regarding translational research.

Chapter 1

Do biomedical models model?

> "A model is [...] almost anything from a naked blonde to a quadratic equation" (Goodman 1968, p.171)

Biologists use the word 'model' in a wide variety of ways. The expression 'a model of oxidative phosphorylation' may be used to designate things as heterogeneous as an engineered bacterial strain or a textbook cartoon of biological processes. While this is surely more than just a linguistic accident, the nature and extent of the resemblance between these different usages is not entirely clear: living, material models are in many ways unlike abstract models. In this chapter, I will be concerned with the question of whether or in what sense they can be called models. I begin with the hypothesis that there is a philosophical sense of 'model' which fits both abstract models and biomedical models, and will not reject this hypothesis before having made a serious effort at finding such an account.

Section 1.1 starts with a brief survey of the philosophical discussion on scientific models, in order to highlight some of the most important features generally associated with models and modeling. Some kind of surrogacy (or indirectness) seems to be central to the notion of model, although its precise meaning remains elusive. The rest of the chapter then attempts to understand how and to what extent these features apply to biomedical models. I will argue that the notion of model, in order to be applicable to biomedical models, ultimately collapses with that of experimental system (section 1.4). This conclusion is reached by discussing two problems in the application of the concept of model.

The first problem is the distinction between models and non-models (or modeling and non-modeling), and the meaning of surrogacy in the context of biomedical models (sec-

tion 1.2). The notion of model is generally opposed to that of an instance, or sample, and therefore relies on pre-established kinds. In the life sciences, it turns out to be very problematic to distinguish between them, and consequently between surrogacy and direct experimentation. As a consequence, one is bound to accept either that biomedical models are not models, or that most of biomedical research – including clinical research – is about modeling.

Assuming we are able to distinguish things that are models from things that are not, this still presupposes a partitioning of research systems into things which may, or may not, bear this ascription. In section 1.3, I discuss model individuation (identifying the boundaries of a model), and propose that models can only be individuated through the robustness of their modeling relationship: two elements are part of the same model insofar as their mapping with elements of the target system are interdependent. This again makes the notion, as applied to biomedical models, co-extensive with experimental systems.

More broadly, the aim of this chapter is to lay down, in the context of philosophical discourse, a series of observations which will then be applied to the analysis of more concrete cases in chapters 3 to 5. Among these is the claim that although biomedical models do share relevant similarities with other scientific models, the vagueness of the concept of model in this context prevents it from being philosophically useful. For this reason, it ought to be replaced by more fine-grained categories, part of which I try to spell out in Chapter 3.

1.1 Philosophical accounts of scientific models

I will briefly go through different accounts of models in order to put the key issues on the table. Note that my aim, here, is not to argue for or against these accounts, but to understand in what sense biomedical models can be said to be models. To this end, informative accounts of models are those which include biomedical models and exclude most things which scientists would not call models. This does not mean that other accounts are wrong: they may well delineate a useful subset of what scientists call models, and might even be a more useful category than scientists' usage of the term. My interest, however, is in biomedical models as understood in section 0.1). In the end, whether we decide to

call these 'models' or some other name is philosophically irrelevant – the notion is but a linguistically motivated starting point in the discussion of what biomedical models are.

1.1.1 Models as mediators

Models have received a lot of attention from philosophers of science throughout the last century. Despite this intense effort (or perhaps because of it), there is little consensus as to a clear account or definition of what scientific models are (Frigg and Hartmann 2012). Probably the most influential accounts have been logico-mathematical: Patrick Suppes' set-theoretical account (Suppes 1960), or the state-space approaches inspired from it (e.g. Suppe 1974). These will not be discussed here, for the main reason that their formal precision is largely incompatible with their broader application to most of what scientists call models, let alone material models such as biomedical models. Attempts to reinterpret notions such as isomorphism in ways weak and flexible enough to fit these (see for instance da Costa and French 2003) have mostly deprived them of their analytical bite. Nevertheless, a basic insight brought by these accounts[1] can be translated into less formal contexts, namely the idea that models are a sort of interface between theories (or knowledge more broadly conceived), on the one hand, and data or phenomena on the other. Depending on the account and context, this can mean providing an interpretation for a theory, putting reality in parameters that can be handled by the theory, making a theory applicable to real contexts, etc.

In this context, the definitional challenge has been to construe models in a way that would successfully distinguish them both from data/phenomena on the one hand, and from theories on the other. As we will see throughout this chapter, the challenge is different in the context of biomedical models, where a mouse model is seldom mistaken for a theory. Nevertheless, the general idea that models mediate (both ways) between reality and our knowledge of it seems intuitively correct in both contexts, although in want of further characterization.

This idea has been picked up for instance by Nancy Cartwright (1983), who sees models as making the link between prepared description of phenomena and theoretical representa-

[1] As well as other traditions – see for instance the accounts discussed in Gayon 2006.

tions, or by Ronald N. Giere (2004) who sees them as mediating between 'principles' and the world. The idea is also central to the 'mediator view' of models (Morgan and Morrison 1999), with the important specification that, according to these authors, models can mediate between theories and the world precisely because they are autonomous from both data and theories. They are "outside the theory-world axis" (Morgan and Morrison 1999, p.18), which also means that they are outside of truth considerations (see also Keller 2000). Rather, models are to be understood as *tools* for theory construction and application. They allow the exploration of "implications of theories in concrete situations" (Morgan and Morrison 1999, p.18), and can also be used "directly as an instrument for experiment" (Morgan and Morrison 1999, p.21), either as surrogates or as measuring devices.

Although models are tools, not all tools are models:

> "The critical difference between a simple tool, and a tool of investigation is that the latter involves some form of representation: models typically represent either some aspect of the world, or some aspect of our theories about the world, or both at once." (Morgan and Morrison 1999, p.11)

On this account, therefore, models are tools in theory construction/application through some form of representation[2]. The notion of representation is common to many accounts of models; unfortunately, philosophical accounts of scientific representation are even more numerous and heterogeneous than accounts of models. The next two sections will discuss the meaning of representation relevant for the present purposes.

1.1.2 Models as representations

"The plain fact is that a picture, to represent an object, must be a symbol for it, stand for it" (Goodman 1968, p.5) *Standing for* something is necessarily standing for it *somewhere*, in some context. In other words, representation is a functional rather than a structural relation: it is not simply a relation between the intrinsic properties of the model (that doing the representation) and those of the target system (that being represented), but rather a relation between the two entities within a larger representational system (what Goodman

[2] Bailer-Jones (2002) reaches a similar, although somewhat more specific definition: "A model is an interpretative description of a phenomenon that facilitates access to that phenomenon." (Bailer-Jones 2002, p.108). She adds that "[t]his access can be perceptual as well as intellectual" (Bailer-Jones 2002, p.108-109), but the omission of experimental access and the focus on models being *descriptions* prevents experimental systems such as animal models from qualifying as models.

would call a language). Giere (2004) therefore writes that instead of being considered as a dyadic relationship, representation "if thought of as a relationship at all, should have several more places". He proposes a tentative formula in this direction:

> "S uses X to represent W for purposes P." (Giere 2004, p.743)

X and W are respectively the model and the target systems. The addition of a subject or community (S) highlights the fact that nothing is a representation independently of representing agents, and the presence of purposes (P) emphasizes that there is no neutral representation and that representations are always with an eye toward their use (Kitcher 2001). However, this leaves open the question of how P impacts on the representation. Following Goodman (1968), Bas van Fraassen (2008) has suggested a slightly different formulation which makes this relation perhaps more explicit:

> "Z uses X to depict Y as F" (van Fraassen 2008, p.21)

The 'as', here, is to be understood as a sort of predication (Goodman 1968). The most obvious example (given by van Fraassen) is that of a caricature, of which we would say that it depicts a famous politician *as* draconian. In a scientific context, one could say that Bohr used the solar system to depict the atom as objects in orbit around an attracting body. In doing so, he allowed agents to apply some previous scheme of reasoning to atoms, but it also distorted some aspects of the atom (to start with, atoms are considerably smaller, and planets do not repel each others). Although he cannot formally prove it, van Fraassen contends that in every such representation, there are necessarily some other aspects of Y that are not (or are inaccurately) depicted.

Thinking of representation *as* can help making the relation between the representation and its purpose clearer. In the case of the caricature, the representation has the purpose of conveying a political message, namely the claim that the politician is oppressive. A failure to carry this message is a failure of representation. This is to be distinguished from successfully carrying a message whose content is false, which would be a misrepresentation. Success or failure of representation is independent of whether Y is accurately described as F, and instead depends on the relationship between X and Y (X must be understood to represent Y) as well as on that between X and F (X must be understood to depict Y as

F). Importantly, both of these relationships are functions of the system of representation (or language). For instance, an artist using an aureola around the head of the politician to represent the politician as draconian would most likely fail to convey his or her message, because aureole have a quite different meaning in our system of representation.

The political context offers another example. Representative democracies can be said to use elected representatives to represent populations as aggregated interests. In this context, a successful representation is a recognized representation (the representative is perceived as representing the aggregated interests of the population), generally as a consequence of transparent elections (and hence depends on the electoral system of choice). However, even recognized representatives may be said to misrepresent the population, if they represent the population as aggregated interests which it does not hold.

It is therefore important to distinguish the relation of representation from its empirical adequacy: virtually anything can be made to represent anything else, depending on the system or language (think of children's games). However, not everything can do so in an informative way. A failure of representation is a communication problem, while a misrepresentation is problem of empirical inadequacy. Note, however, that this is a very special kind of empirical adequacy, for given purposes: political representatives can be in many respects unlike the population they represent, and yet they do not misrepresent the population unless they actually misrepresent its aggregated interests. In Bohr's model (or, more precisely, its reception), the fact that the solar system is orders of magnitude larger than the atom does not misrepresent the atom, because the kind of system *as* which the models represents the atom does not include size. This shows the importance of practical purposes in establishing the value of a model or representation, but it also highlights that the relation of adequacy is not between the representation and the represented, but between the represented (Y) and that *as* which it is represented (F). In the case of the caricature, the adequacy of the representation depends on whether the politician (Y) is draconian (F) or not, a question which is independent from what is doing the representing (X).

1.1.3 Similarity and adequacy of representation

van Fraassen argues that "Not all, but certainly many forms of representation do trade on likeness, likeness in some respects, *selective likeness*." van Fraassen (2008, p.7). The 'not all' seems to be made necessary by acts of arbitrary denotation: in games (which includes much of social interactions), the players can agree that a certain item represents another independently of their similarity. For instance, we might say of any piece of wood on a map that it represents the position of a ship. Considering that "[a]nything is similar to anything else in countless respects" Giere (2004, p.747), one might argue that the piece of wood is actually similar to the ship (both are made of wood), but the point is that it could equally not be – that it is not by virtue of this similarity that it is representing.

The relation of representation is therefore independent of resemblance (Goodman 1968, p.5, Suarez 2004), at least of resemblance between the source and the target of representation. However, this does not necessarily mean that the adequacy of a representation (for given aims) is independent from resemblance, i.e. resemblance might be what distinguishes representation from misrepresentation. The caricature is adequate if the politician resembles the idea of being draconian. This, however, is already an unnecessary commitment to realism; furthermore, it leaves open the question of how similar Y and F must be. Consider the purpose of the piece of wood on the map: it serves to show the location of the ship – it depicts the ship as being located in some specific ways with respect to other elements of the map. Whether this location is precise enough depends on what it will be used for, on the actions it will lead agents to take. Adopting a realist interpretation of the adequacy would therefore imply an important metaphysical commitment, and yet would not bring any advantage, for it would not rid us of the necessity of considering the pragmatic dimension.

More importantly, similarities may impair the purpose of the representation. van Fraassen's idea of 'selective likeness' suggests a sort of sufficiency[3]: the model need not be similar to the target system in all respects, but it is sufficient that it is similar to it in some relevant respects (there is a widespread agreement on this issue – see for instance Suárez 2010, p.95 and Godfrey-Smith 2006). An implication of this view is that although lesser similarity is not necessarily worse (it is neutral if the dissimilarities are in irrelevant aspects), greater

[3] For instance, he writes of "sufficiently accurate resemblance in all relevant respects for the purpose at hand." (van Fraassen 2008, p.49)

similarity is necessarily neutral or better. However, dissimilarities between the model and the target system may foster rather than impair the purposes of the representation. A map of the subway system does not need to represent every aspect of it, indeed it would be problematic if it did, for we would have a much harder time figuring our way. Likewise, there are different ways in which portraits can represent the same subject. Between a photograph and a caricatured version of the same photograph, the photograph carries more (and more accurate) information about the subject. However, when it comes to face recognition, caricatures are more efficient precisely because they emphasize features that are specific to the individual (Rhodes et al. 1987, Mauro and Kubovy 1992). Hence if the representation is used for purposes of identification, a caricature is a better representation of the subject. This example shows that the adequacy of a model is not reducible to features of the model and target systems, but has to take much more into account. Indeed, it is important that in the case of caricatures, the dissimilarities are not informative about the subject itself (in which case they would simply represent similarities of another kind), but about the relation of the subject to the rest of the population. A caricature, therefore, does not only stand in a relation with the person it represents, but it situates a person with respect to human facial traits – *it locates the person on a special kind map*. It is, in other words, theory-laden (although there is no formal theory), which is not a weakness, but the very condition of possibility of measurement.

This is what I want to emphasize here, and which will be further developed in chapters 4 and 5: that *as* which Y is represented (F) should most often be conceived not as a property that Y is taken to have, but as a position in a space of representation. One consequence of this is that one cannot assess a model independently of the rest of the representational system within which the modeling operates, and which generally include other models.

1.1.4 Modeling as surrogacy or indirect access

The importance of the practical application of the representation (and hence of the model) in the establishment of a relation between it and the target, as well as in the justification of this relation, has lead some philosophers to tackle the question the other way around. Instead of trying to define what kind of relation there must be between the representing

and the represented in order to know when there is representation, some philosophers have proposed to ask what job representation is doing in order to know the kind of relation which can do the job. In this line, Mauricio Suàrez suggested to "take surrogate reasoning to be the primary function of scientific representation" (Suarez 2004, p.770; see also Swoyer 1991). Following a similar reasoning, Michael Weisberg also proposed to take surrogacy – or indirectness of access – as the defining features of the activity of modeling (see also Godfrey-Smith 2006) :

> "Modeling, I will argue, is the indirect theoretical investigation of a real world phenomenon using a model. This happens in three stages. In the first stage, a theorist constructs a model. In the second, she analyzes, refines, and further articulates the properties and dynamics of the model. Finally, in the third stage, she assesses the relationship between the model and the world if such an assessment is appropriate. If the model is sufficiently similar to the world, then the analysis of the model is also, indirectly, an analysis of the properties of the real-world phenomenon." (Weisberg 2007, p.209)

Weisberg distinguishes modeling from what he calls "abstract direct representation" (ADR). His main examples, respectively of modeling and ADR, are the Lotka-Volterra model of predator-prey interactions, and Mendeleev's periodic table. The former is the classical example of a mathematical model in biology[4], and can be represented (following Vito Volterra's description) as the following pair of differential equations:

$$\frac{dN_1}{dt} = (\epsilon_1 - \gamma_1 N_2)N_1$$

$$\frac{dN_2}{dt} = (-\epsilon_2 + \gamma_2 N_1)N_2$$

Where N_1 and N_2 represent, respectively, the sizes of prey and predator populations; ϵ_1 represents the prey's growth rate and ϵ_2 the predators death rate; and γ_1 and γ_2 represent respectively the predator's voracity and the prey's defense capacities (see Volterra (1926, 1928); for an interesting philosophical analysis of Volterra's and D'Ancona's original work, see Scholl and Räz (2012)).

The model made some largely correct predictions such as oscillations in population sizes and, more importantly, explained the observed fact that heavy fishing tended to favor

[4] The Lotka-Volterra model has its name from its parallel development by Alfred J. Lotka especially in the context of organic chemistry, and slightly later by Volterra in the context here discussed.

the prey population, which was D'Ancona's original puzzle. Weisberg argues that these equations

> "were not direct representations of any real system. It was only in virtue of the similarity between the models he had characterized and real populations of fish in the Adriatic that Volterra could answer D'Ancona's query." (Weisberg 2007, p.216)

I believe the best way to make sense of Weisberg's position is following Peter Godfrey-Smith's view according to which the models are not (or not limited to) the mathematical equations, but they are the "imagined concrete things" (Godfrey-Smith 2006, p.736) specified by the equations. In the case of Lotka-Volterra's model, this would be imaginary populations (of entities) whose sizes are related in deterministic ways through the relationships described by the equations.

Dmitri Mendeleev's periodic table of elements also involves important abstractions: of all properties of elements, only some were selected for the theoretical representation. However, Weisberg argues that Mendeleev "worked directly with abstractions from data" (Weisberg 2007, p.215) to create his periodic table of elements, in other words that there was no surrogacy involved.

Weisberg notes that the distinction "is about the practice, not the products of theorizing" (Weisberg 2007, p.228). Once the model is successfully compared to the phenomenon, it is possible that it is not different in nature from the outcome of ADR. The difference, according to Weisberg, is in the epistemic stance and strategy of the modeler (see also Scholl and Räz 2012) – that in modeling, operations are performed on the model *before* its adequacy to the real phenomena is assessed.

On closer inspection, however, Weisberg's distinction is far from clear-cut. He writes that Volterra "did not arrive at these model populations by abstracting away properties of real fish, he constructed these model populations by stipulating certain properties." (Weisberg 2007, p.210) However, this first step of the modeling already depends on direct abstract representation of the little one knows of the phenomenon. Indeed, it is telling that the variables of the model were from the start interpreted in causally suggestive ways, for instance the 'voracity' of the predators. Clearly, this parameter was not, as Weisberg suggests, a mathematical artefact later interpreted in terms of properties of the real system,

but it was interpreted from the very beginning. Nor did Volterra come up with equations of this particular form by chance. The basic entities and relationships of the model were direct abstractions from experience: the observation of fish eating other fish, and of fish reproducing or fish populations growing. The 'imagined concrete things' of the model system and the relationships between them were direct abstractions from the phenomenon. Because these abstractions were known to be adequate to the phenomena before the construction of the model, Weisberg's first stage is not dissociated from the phenomena or from issues of empirical adequacy.

If the first stage cannot distinguish models from ADR, perhaps one should turn to the later stages of modeling, especially to the fact that in modeling, one first "articulates the properties and dynamics of the model" before comparing these to the actual system. However, cases of ADR generally involve some amount of surrogative reasoning. It seems very likely that Mendeleev tried several organizational schemes before reaching his final representation, and each time he confronted this schemes with what he knew of the elements. What we know for sure is that in 1869, as Weisberg notes, Mendeleev predicted the existence of three new elements on the basis of "the chemical trends encoded on the Table" (Weisberg 2007, p.213). Clearly, these were elements of the representation which were then compared to their hypothesized analogues in the world. The problem is that most theories, inasmuch as they are generators of testable hypotheses, involve surrogate reasoning on the theoretical system followed by confrontation with reality.

Note that the arguments presented in this section are not *per se* reasons to abandon Weisberg's account. First, it might simply be that Weisberg's choice of the Lotka-Volterra model as a paradigmatic case of modeling was unfortunate. Indeed, there are cases of mathematical modeling where at least components of the model are purely theoretical until the model's adequacy to a phenomenon is assessed. For instance, Duhem (1906) complained of Maxwell's equations that they contained parameters that had no interpretation. Limiting modeling to practices of model-fitting would however lose most of what scientists call modeling, even more so in biology. Alternatively, one might accept that the difference between modeling and ADR is one of emphasis: of whether direct abstraction or surrogate reasoning is the most predominant in the scientist's activity.

Weisberg's account of models has the important downside of being incompatible with the theory-model distinction in the semantic account of the structure of scientific theories (see Suppe 1974), and it now seems that his account of modeling is unable to fully replace this distinction. However, this problem of distinguishing models from theories is due to the fact that many scientific models are linguistic representations (in the broad sense of Goodman 1968, which includes graphical and mathematical representations) created for the double-purpose of being manipulated and of encoding some of our knowledge. Indeed, when a biologist draws a cartoon of a molecular mechanism, it is not uncommon to say that "this is what we know of the mechanism". This implies that the models have to be produced, or at least arranged by the cognitive agent(s). Although biomedical models are increasingly modified, crafted for specific purposes or standardized (Kohler 1991), the sense in which they can be said to encode or represent our knowledge is very limited[5]. Perhaps, then, the difficulties of Weisberg's account disappear in the case of biomedical models, because they are not knowledge representations, as the product of ADR is. A mouse is not a theory, and neither is it a human being. Yet it is used to study human biology, and therefore it must be a model. However, as section 1.2 will show, even in this context the distinction between models and non-models is not so straightforward.

The notion of surrogacy seems essential to the definition of biomedical model adopted by the National Research Council's (NRC) committee on the topic: "A biomedical model is a surrogate for a human being, or a human biologic system" (ILAR and NRC 1998, p.10). It is a central aspect of some of the most promising accounts of modeling and represents a sort of implicit consensus in the discussion of biomedical models (LaFollette and Shanks 1996, Godfrey-Smith 2006, Bolker 2009, Shanks and Greek 2009, Greek and Shanks 2011, Piotrowska 2012). Earlier accounts of modeling (section 1.1.1) also involve a sort of *indirectness*: for a model to be a mediator, it has to stand in some way between the world and our theoretical understanding of it, or provide a prepared (Cartwright 1983) or facilitated (Bailer-Jones 2002) version of it. While these are certainly relevant resemblances,

[5] In a certain way, some experimental organisms can be interpreted in this way: a mouse genetically engineered to develop cancer contains oncogenic lesions which can be said to be material inscriptions of our knowledge of carcinogenesis – "thing knowledge" (Baird 2003). But this would be akin to saying that a building is an inscription of our architectural knowledge: while it is true from some perspective, the representation was not made for this purpose and is a rather poor way to store and transmit knowledge. In contrast, the textbook model of a mechanism has this function.

the exact meaning of this indirectness is rather elusive, and I will show throughout the chapter that the problem is even more acute in the context of biomedical models.

1.1.5 Model of, model for

As different authors have noted, the word model is used in two different ways in science: models *of*, and models *for*[6]. Philosophical attention to the distinction is generally attributed to Evelyn Fox Keller (Keller 2000 and Keller 2002), but she uses it without explicitly defining these expressions. Nor is her usage trivial, for in fact the expressions are used with reference to the same things[7] – although not necessarily to denote the same aspects of them. This suggests that models are always to some extent both models-for and models-of, and that the expressions merely *stress* one aspect or the other. This is the position of anthropologist Clifford Geertz when tracing the same distinction:

> "The term 'model' has, however, two senses – an 'of' sense and a 'for' sense – and though these are but aspects of the same basic concept they are very much worth distinguishing for analytic purposes. In the first, what is stressed is the manipulation of symbol structures so as to bring them, more or less closely, into parallel with the pre-established nonsymbolic system, as when we grasp how dams work by developing a theory of hydraulics or constructing a flow chart. The theory or chart models physical relationships in such a way – that is, by expressing their structure in synoptic form – as to render them apprehensible; it is a model *of* 'reality'. In the second, what is stressed is the manipulation of the nonsymbolic systems in terms of the relationships expressed in the symbolic, as when we construction a dam according to the specifications implied in a hydraulic theory or the conclusions drawn from a flow chart. Here, the theory is a model under whose guidance physical relationships are organized: it is a model *for* 'reality'." (Geertz 2000, p.93)

A first way to interpret Geertz is to say that a model is a model 'of' when we are building it ("the manipulation of symbol structures so as *to bring them, more or less closely, into parallel* with the pre-established nonsymbolic system"), and a model 'for' when we are using

[6] It is important not to confuse this distinction with the related (but distinct) distinction by Mary Morgan between 'representative for' and 'representative of': "While inferring to other objects in the 'representative of' case relies on establishing representativeness in the same sense as the sample/population relation (that is, they are of the same case), inferring in the 'representative for' case depends on establishing similarity relations between two different organisms" (Morgan 2003, p.228) Unless understood as a subjective perspective on the model, Morgan's distinction is highly problematic in biology in that it requires natural kinds (see section 1.2.1).

[7] For instance, Alan Turing's reaction-diffusion model is said to be "a model of embryogenesis" (Keller 2002, p.6, p.12, p.285), "a model of morphogenesis" (Keller 2002, p.319, note 61), a "model for morphogenesis" (Keller 2002, p.96) or a "model for embryogenesis" (Keller 2002, p.189).

it for something ("the manipulation of the nonsymbolic systems in terms of the relationships expressed in the symbolic"). However, most often in biology these do not represent distinct moments, and the common usage of the expression suggests that a model does not cease to be a 'model of' something merely because its construction is over. 'Model for' is 'for doing', for concrete interventions, such as building a dam, but this does not preclude it from also being a model of (aligning with) a dam.

While 'for' is suggestive of action, 'of' (which is often followed by a noun) is suggestive of representation. Of course representing is itself an action (like interpreting, explaining, describing, thinking, etc.)[8], and it can be done in concrete ways (e.g. on a piece of paper). In fact Geertz writes of 'manipulation of symbol structures', and material symbols are not precluded. The division is not between cognitive and material, but between symbolic and non-symbolic. It is, in other words, about whether some things stand for something else. Importantly, however, it is not the model system that stands for the target system, but the model as a symbol structure which 'aligns' with the target structure: it is elements of the model that stand for elements of the target. This is an important distinction to be made, for the fact that features of the model stand for features of the target does not imply that the model itself stands for the target – an assumption which I will discuss in chapters 2 and 3.

In the context of a discussion on model organisms, Jean Gayon (2006) also takes the 'model of' aspect to be related to the scope, or to the represented. He writes that

> if a geneticist presents a 'model for (the study of) genetic regulation rather than a model of genetic regulation, it is because he knows that there is not a unique model for genetic regulation, and that he believes that the model he presents has a heuristic value.[9]

This point is extremely important for model organisms. The heuristic role of model organisms is not limited to the generation of hypotheses through the suggestion that something in our organism of interest 'might be like in the model organism'. Even when the hypothesis

[8] Jean Gayon (2006) generally interprets Keller's usage of 'model for' as meaning 'for the study of', an interpretation which I find consistent with Keller's texts, but which leaves open the nature of this study. It should not come as a surprise to see the distinction discussed in a paper which "does not begin by presupposing an a priori divide between theory and practice" (Keller 2000, p.S73).

[9] "Pour reprendre l'exemple pris par Keller, si un biologiste présente un 'modèle pour [l'étude de] la régulation génétique' (*model for genetic regulation*) plutôt qu'un modèle de la régulation génétique, c'est parce qu'il sait bien qu'il n'y a pas un unique modèle de la régulation génétique, et parce qu'il estime que le modèle qu'il présente a un intérêt heuristique." (Gayon 2006, p.27)

is refuted – when, to take a simple example, an intervention turns out to have a different outcome on mice and on humans – the knowledge we have about the model organism offers strong suggestions as to how to go about explaining this disanalogy and hence understanding the target system. In other words, the mouse can be a good model *for* the study of a human phenomena even without being a good model *of* this phenomena.

This being said, a model cannot be a good model *for* anything unless it is also a model *of*: if a model is informative about something, it is because it has some sort of similarity with that thing, however abstract. Even purely statistical models without underlying mechanistic insights must have some similarity with what they model. Marcel Weber (personal communication) has therefore suggested that the distinction between models of and models for hinges on whether or not there is an explicit mapping function between the model and the target. A model is a model *of* if there is a mapping function indicating what it is about the model that is representing what in the target (what features of the model stand for what in the target). In Weisberg's most recent account of modeling, this is specified in his first step of modeling, what he calls assignment: "Assignments are explicit specifications of how parts of real or imagined target systems are to be mapped onto parts of the model" (Weisberg 2013, p.39; contrary to what the quote suggests, Weisberg considers assignments to be most often implicit – see Weisberg 2013, p.40). In contrast, a model *for* assumes that there must be some kind of mapping function, but makes no commitment about its nature or breadth.

Similarity approaches to modeling (sections 1.1.2-1.1.3) have concentrated on its model-of aspect, while functional approaches (section 1.1.4) have focused on its model-for aspect. They are two sides of the same coin.

To sum up, therefore, models can be understood as tools in theory construction and application (section 1.1.1), allowing surrogative reasoning or indirect access (section 1.1.4) by virtue of a relation of representation (section 1.1.2) involving selective (and highly context-dependent) similarity and idealization between the model and the target system (section 1.1.3). I do not wish to claim that these features are necessary or sufficient for something to be a model or be used in modeling, nor is such a definition needed for what follows. The aim was instead to get an idea of the important features shared by most models or cases of

modeling, in order to guide their interrogation in the context of biomedical models. In the remainder of this chapter, I will try to see to what extent biomedical models (or modeling) have these features, and translate to this context the observations just made.

1.2 Surrogacy in biomedical models

> "The primary meaning of the term model in experimental biology is an organism, an organism that can be taken to represent (that is, stand in for) a class of organisms." (Keller 2002, p.115)

The first thing one is confronted with when comparing models in biology with other usages of the term is their materiality. However, it is not materiality *per se* which is relevant: it does not seem to make any important difference that Watson and Crick constructed a cardboard model of DNA instead of simply imagining it and describing it with words. Materiality is relevant only insofar as it is used to provide empirical input, rather than only conceptual import (see Leonelli (2007)). What is particularly distinctive of concrete biological models is that they somehow embody the phenomenon of interest (Griesemer 1990): Drosophila was not simply a model for the study of genetic transmission, but it was also an *instance* of genetic transmission. This creates a tension with the philosophical account of modeling: if models are characterized by surrogacy or indirectness of access, and if biomedical models literally contain the phenomenon they model, then this can hardly qualify as a surrogate[10].

In a sense, this is true of all models: playing around with the Lotka-Volterra model informs us about oscillations in a system of differential equations (or, following Godfrey-Smith, in the imaginary concrete populations they describe), at the same time as it informs us about variation in adriatic fish populations. The problem is not with something being simultaneously a sample and a model, but rather in making the difference between the two. In the case of the Lotka-Volterra model, the model *is* a system of differential equations (or the imaginary populations), but it is not a fish population. For this reason, it can be said

[10] Goodman also noted that "In many cases, a model is an exemplar or instance of what it models" (Goodman 1968, p.171), and concluded that the label "might well be dispensed with in all these cases in favor of less ambiguous and more informative terms, and reserved for cases where the symbol is neither an instance nor a verbal or mathematical description: the ship model, the miniature bulldozer, [...]" (Goodman 1968, p.172). Goodman unfortunately gives no clue as to why the ship model is not a ship: what is the size at which a boat stops being a boat?

to be simultaneously an instance of the first and a model of the latter. The question is whether one can draw the same distinction in biomedical models.

1.2.1 Cutting lineages at the joints

In order for there being surrogative or indirect means of studying something, there has to be at least in principle non-surrogative means, which I will call 'direct experimentation'. Trying to define what this direct experimentation is turns out to be problematic, especially in the life sciences. The reason is that while the physicist can claim that he is studying 'electrons in general' while performing experiments on a very small and non-random subset of electrons, the biologist dealing with supra-molecular entities cannot. Electrons are natural kinds in the strongest sense, or at the very least can be considered as such for most scientific purposes. In the life sciences, however, similarity between members of a class is generally (if not always) a matter of degree.

When biologists conduct experimentation on a species – and even if it is in its natural environment – they seldom follow every individual in the population, and do not study future or past individuals. Likewise, when we experiment 'directly' on humans, a sample of individuals stands for the rest of the population. This leaves us with two choices: either we accept that the vast majority of experimental biology and medicine are cases of modeling, or we must distinguish samples from models. The most obvious strategy to ground this distinction would be to argue that generalization is an inference from tokens to type, whereas extrapolation is between types. This presupposes an organization of biological entities into pre-existing types: if the type we consider is the human species, then the mouse is used as a model rather than a sample; however, if we take mammals to be the type, then it is a sample. Species are not natural kinds, or at least not in the way electrons are. Because species do not have an essence[11], morphological or phenetic approaches must rely on similarity, and there is no objective choice among the indefinite number of measures of similarity. Lineage-based concepts, on the other hand, are at trouble singling out the right granularity for species, as well as where a species start. Finally, it is unclear why, for

[11] It has been argued that species can be natural kinds under some accounts of natural kinds, such as the homeostatic property clusters account (Boyd 1999). In most cases, however, this does not solve the problem of granularity: they nevertheless remain one out of many 'natural' and useful ways of classifying organisms.

the purpose of extrapolation, one should adopt the the biological species concept (which defines species by reproductive isolation)[12]. The only reason why reproductive isolation would matter to issues of extrapolation is that inbreeding is a good proxy for similarity. Once more, however, similarity comes in degrees and in many different flavours. Members of a same species also differ in many relevant respects, often making one individual a bad surrogate for another.

It is therefore unclear why, for the purpose of generalization, one should cut at species rather than at other (broader or narrower) *taxa*. As a matter of fact, the NRC report mentioned earlier tends to draw the line not between humans and other species, but between primates and other species:

> "Studying human health involves two general experimental approaches: examining human or primate cells, tissues, and organs that constitute relatively direct models of human disease; and using a variety of model systems that offer special features and advantages that are not available for study in human beings or primates but can be applied to human health issues." (ILAR and NRC 1998, p.6)

There are many mysteries in this way to categorize things, but one thing is clear: extrapolating to a specific mouse strain, to rodents or to mammals is a matter of degree, and therefore something else must determine where to cut. The reason is simple: whether within or between species, it is always phylogenetic proximity, as a proxy for similarity, that makes it possible to learn about individuals through other individuals. I am common to other human beings first and foremost because of our closely shared ancestry. This can be seen at work in the necessity to take ethnic groups into account in contemporary clinical trials (see for instance Bloche 2004). Ideally, a sample should represent individuals of each major ancestry of a population, and for an individual of a different genetic make-up, results have to be interpreted with care. In fact, for specific purposes animal models may be a better surrogate for a given patient than most other humans. For most relevant purposes, a genetically engineered mouse model of cystic fibrosis will be more similar to a human patient suffering from the disease than a healthy human would be.

[12] In their famous critique of animal models, Shanks and Greek (2009) adopt a cohesion-based approach to species, relying on what they call "intrinsic cohesion mechanisms" (Shanks and Greek 2009, p.150). While such approaches may in some respects represent an improvement over the biological species concept, the points raised here apply equally to them.

From the point of view of generalizing results, the distinction between clinical and animal experiments is one of degree rather than nature – which does not mean, obviously, that it is irrelevant.

1.2.2 Samples in mosaic individuals

But let us probe this point further to see its full scope. There are cases in medical practice where one makes a measurement on a patient in order to learn something about the patient – surely this should be 'direct' in a way that modeling is not. While this may be true of some measurements, most (especially in molecular medicine) involve experimental manipulations that make this directness problematic.

To begin with, the instability of cells is exacerbated by laboratory manipulations: cells adapt to their culture conditions, in both the physiological and evolutionary senses. Culture affects their phenotype, behaviour, and expression profile (see for instance Creighton et al. 2003), and might even result in the accumulation of genetic alterations, potentially leading to a selection for life in a dish. In this context, the similarity of cultured cells to the cells in the host's body is not taken for granted, but an ideal objective which is often in tension with the technique itself[13]. Extrapolating from cultured cells is not any simpler than extrapolating from an organism of another species.

Even with minimal manipulations, the cells of a sample is often from the start different from what was sampled. Consider the following example. There are many different classifications of breast cancer that cut across each other, and some of the most important categories are determined by the presence or absence of certain molecular markers. For instance, HER2-positive breast cancers (cancer with high expression of the protein HER2) are generally much more aggressive than HER2-negative, and yet are also very responsive to a potent drug (trastuzumab, often considered a successful example of targeted therapy). It is therefore important to test for this marker in invasive breast cancers. However, the test (be it for a mutation or for expression) is performed on a biopsy – on a small piece of the tumour removed from the patient. Tumours are heterogeneous tissues, and as recent

[13] In a NRC meeting on the respective value of *in vitro* and *in vivo* models, Mary Dawson noted of tissue culture that "[t]he object of this type of culture is to keep the cultured material as near as possible to its in vivo appearance and function and at the same time have improved access to it." (Institute of Laboratory Animal Resources (ILAR) 1977, p.188)

studies have shown, the variation between two biopsies taken from the same tumour of the same patient (on the same day) can be considerable (Gerlinger et al. 2012, Shibata 2012). Strictly speaking, the cells on which the test is performed are not those about which we wish to learn: cells from the biopsy are not in the patient anymore, and therefore are clinically irrelevant. However, they are believed to be informative about the remaining cells of the tumours.

Whether this information is credible (whether the sample is a good model) depends on a number of factors. Cancer is characterized by genetic (and epigenetic) instability, which makes this divergence relevant. However, there is a growing body of evidence suggesting the importance of somatic variations in normal development as well (Gupta and Poss 2012, Singer et al. 2010, Jablonka 1996). In this context, and given the potential issue of inadequate representation, tissue samples could well be considered as models.

The issue is even more important in the context of cellular reprogramming, which allows the conversion of cells from one cell state to another. I will extensively discuss the applications of this technology for disease modeling in Chapter 5, and for the moment I would simply like to note two things. Until recently (Rais et al. 2013), cellular reprogramming to a pluripotent state was a particularly inefficient process, which meant that only a few cells will make it out of thousands. As a consequence, it carried the risk of selecting for rare mutations in the cell population that facilitate the cell fate transition. However, recent studies have shown that most of the genetic variation is reprogrammed cells are instead due to background mutations present in low frequency in the tissue of origin (Young and Goldstein 2012). From an evolutionary point of view, it might be reasonable to say of someone that he/she has a genome or genotype, but from the point of view of health and development, we are all mosaic individuals: our body contains mixed populations of cells with different genotypes. Nevertheless, just like the cells of any primate ultimately come from the same ancestor as my own cells, so do the reprogrammed cells ultimately derive from the same (embryonic cells) as the neurons of the patient. In both cases, it is shared genetic lineage (be it ontogenic or phylogenic) which, *ceteris paribus*, warrants an assumption of high genetic similarity. Admittedly, the fact that somatic cells reproduce asexually makes an important quantitative difference, but the point here is that the distinction is one

of degrees.

Blood samples might seem less problematic in this respect than samples of solid tissues, in which cell lineages cluster spatially. Because it is liquid and circulating, blood samples are a more or less random selection of the blood cells, and therefore should accurately represent an average of the whole tissue (or, more exactly, averages over each type of blood cells). Nevertheless, while blood samples might avoid biases due to spatial location, they are nevertheless located in a temporal manner. Indeed, Claude Bernard noted that not only an organism is not entirely comparable to another organism, but it is also not entirely comparable to itself at a different time (Bernard 1865, second part, chapter II §VIII). While these variations might be irrelevant for many purposes, it is highly relevant for instance for research on stress, because the simple act of taking a blood sample changes the composition of the patient's blood. Ultimately, for most purposes *even I am but a model of myself*, for the simple reason that this 'myself' in which I am interested is a construct not reducible to any particular spatio-temporal sample. This is not a trivial point, and I will claim in chapters 4 and 5 that the target of modeling is often (and ought to be) a theoretical construct. But for the moment, my aim was simply to show that this attempt at distinguishing models from samples has important limitations.

1.2.3 Material idealizations

Finally, a last strategy to ground the model/non-model (or model/sample) distinction would be to claim that while samples are simply picked up from a population, models are idealized – they are material idealizations.

Typically, while abstraction is simply leaving out, in the model, certain features of the phenomena, idealization involves ascribing the model features known to be false of the target phenomena. At first glance, this might seem to be highly dependent on a frame of reference. In Newtonian mechanics, for instance, considering objects as punctual masses can be interpreted as abstracting away from the distribution of their mass across space, or as granting the objects a property (being punctual) which they do not have[14]. Instead, I believe that the difference hinges on whether the features are used in the model: punctual masses

[14] Goodman (1955) has convincingly argued that there is no logical way to bar such predicates.

are an idealization because the punctuality is necessary for many features and equations of mechanical models to hold. However, the color of the objects is abstracted away because it enters nowhere in the equations.

Idealization is often pictured as a voluntary distortion, in other words the distortions are in the design of model. While the extent to which material-biological models such as animal models are *crafted* has been criticized (Weber 2005), it cannot be denied that animal models, especially in contemporary biomedical research, involve an increasing degree of modification, from standardization through inbreeding (Rader 2004) to genetically engineered models and humanized models (see for instance Maugeri and Blasimme 2011). I see no problem in viewing biomedical models as material idealizations, but this will not help the distinction we are trying to draw. As physician and philosopher Georges Canguilhem emphasized early on, the objects of most of experimental biology are artefacts (Canguilhem 1965, p.28), and he pointed out that the same can be said of most science to the extent that it studies a new nature, superimposed to the first. Indeed, even clinical trials and clinical encounters are to some extent artificial, and participants in clinical trials are also increasingly being modified for the purposes of research. Very often, for instance for purposes of tracking, radioactive substances are injected into patients. Likewise, diabetes patients in clinical trials often have implants to monitor sugar and insulin levels. In some cases, these modifications are even more dramatic: in a recent study on glioblastoma recurrence, patients had metal clips permanently implanted in their brain, without any therapeutic benefit and solely for the purpose of studying the dynamics and retrospective relevance of different subsets of cancer cells in the expected recurrences of the tumour[15]. Obviously, the extent of crafting is on average greater in so-called models than in patients, but these glioblastoma patients are arguably more intensely modified than many relatively intact animals in biomedical research. For this reason, the amount of crafting involved is a very poor ground for distinguishing models from samples.

Models, however, can be idealized without active and explicit distortion. Animal models, perhaps, are more simple independently of our craft. Indeed, the rationale of model-organism research is often presented as studying 'simpler' organisms is an approach to

[15] The project is headed by Björn Scheffler of the University of Bonn; the results have not been published yet.

complex organisms. While this may seem relatively uncontroversial when comparing unicellular organisms with multicellular organisms, such a claim is much more difficult to justify within, say, higher vertebrates. Complexity is a notoriously difficult notion to operationalize, and there seems to be no good reason to think that one mammal is more complex than another[16]. However, simplicity needs not be cashed out in terms of intrinsic complexity, and can be a pragmatic form of simplicity. Mice are more simple than human beings because they are smaller, breed faster, etc. They are in this sense ideal (economical) patients.

Once more, however, the same applies to clinical trials. Participants in a clinical trial are also idealized in this sense, for they are selected because they are pragmatically simpler. Indeed, the average clinical trial participant is not the average member of the population. This is in part due to exclusion rules, and in part to factors external to the trial procedure. The most simple such pragmatic simplification is to use patients willing to participate. Among the exclusion rules, at least some are clearly aimed at selecting idealized patients, in the sense of pragmatically simpler patients. For instance, the common exclusion of aged patients, or of patients who suffer from other major diseases on top of the one of interest (co-morbidities), are meant to reduce the probability of externalities. Likewise, in some specific fields it is common to exclude participants who are deemed at risk of non-compliance. As in other cases of modeling, epistemic power can be gained at the cost of some representativity.

1.2.4 Abandoning the model/sample distinction in biomedical research

Unless we are ready to accept that clinical sciences are an instance of modeling, thereby undermining the need for a distinction between models and samples in biomedical research, the distinction cannot be established on some absolute ground. However, perhaps it can be contextually defined. Since biomedicine is interested in applying its findings to humans, one might suggest, it identifies *ipso facto* humans as the target group, and excludes other

[16] Shanks and Greek (2009) reject the idea that animal models are inherently simpler: "In this sense, animal models are unlike models found in many other branches of science, where, for example, the model system is usually considered to be a simpler system than the system modeled – an idealization, perhaps, in which needless complications can be ignored (or hidden in a host of *cateris paribus* clauses)." (Shanks and Greek 2009, p.33)

organisms. Members of the target group are therefore samples, whereas organisms outside this group can at best be models. The problem with this argumentative strategy is that on this account, most (though not all) cases normally called 'model organisms' would rather be samples, insofar as they also teach us about themselves. Thomas Hunt Morgan's *Drosophila* would be a sample of organisms whose genetic material is organized in chromosomes. While there might be no problem in saying that *Drosophila* was a sample of eukaryote genetics and a model for human genetics, the fact that humans are also eukaryotes means that experiments on *Drosophila* informs us about humans in two different ways, only one of which would be modeling. If the distinction between them is not simply a matter of how extrapolation is described, but it instead maps genuinely different practices, then perhaps it can be a meaningful distinction. However, I doubt that this is the case, and in Chapter 3 I will instead propose that replacing the notion of model with narrower, more precise concepts can clarify epistemological analysis.

1.3 Animal models, model animals

The first part of this chapter tackled the question of whether something is a model or not. This presupposes a partitioning of systems, i.e. it supposes that one already knows what the candidate models are. However, the issue of model individuation is also not straightforward in biomedical research.

The expression 'animal model', very common in biomedical research, suggests a model having the property of being animal, as opposed to an animal having the property of being a model (as is the case for the expression 'model organism'). In other words, an animal model would be a model which involves (contains) animals. Such a distinction was made for instance by Cameron Shelley (2010), who gives the example of the Porsolt's Forced-Swim Test. This behavioural despair test simply involves a mouse (or other small animal) immersed into a cylinder of water. The mouse first tries to escape the cylinder, but is unable to do so, and after a while it stops and floats passively. Interestingly, giving the mouse antidepressants makes it try *longer* (Porsolt et al. 1977), and the test is therefore known as a model for depression.

There would be much to unpack in this example, but for the moment I simply want to

discuss the following remark by Shelley:

> "Briefly put, many commentators construe animal model to mean the kind or species of animal that is used in a given test. For example, the Porsolt Forced-Swim Test may be called a 'mouse model' of human depression. Indeed, it is a mouse model but the mouse is not the whole of the model. The *entire regime to which the mouse is subjected* comprises part of the model also. That is, the administration of stimulant, the cylinder of water, the starting and stopping conditions of the test, are all part of the model too." (Shelley 2010, p.297, original emphasis)

Shelley is making a very general point, namely the fact that the mouse cannot in itself be a model of (or for) depression, and therefore is not *per se* the model[17]. His example makes it particularly striking because it is clear that the mouse is not the only difference with the phenomena being modeled, and that other elements in the set-up are *standing for* their clinical counterparts in a way that is necessary for the whole system to be a model of/for depression. Indeed, floating passively in water is not the standard clinical description of depression (nor of despair), but an analogical phenomena.

As noted earlier (see especially section 1.1.5), it is not so much (or not only) the model that stands for its target, but features of the model that stand for features of the target. As Giere noted, models "are designed so that elements of the model can be identified with features of the real world." (Giere 2004, p.747) But just as features of the mouse stand for features of a human, features of the experimental setting stand for features of the human situation. The logic of Porsolt's swim test is that being immersed in water and unable to get out of it is supposed to be, for the mouse, what it feels for a human being to be stuck (immersed) in an unpleasant situation. The cylinder full of water stands for the depressing environment. The activity of science, as Hans-Jörg Rheinberger writes, "consists in producing, in a space of representation, material metaphors and metonymies." (Rheinberger 1995a, p.115)

This is not limited to models in psychiatry. Consider the recent publication by Xu et al. (2013), titled "Effects of Perinatal Lipopolysaccharide (LPS) Exposure on the Developing Rat Brain: Modeling the Effect of Maternal Infection on the Developing Human CNS". Their findings are irrelevant for our purposes, and I am only interested in the research

[17] Some commentators seem to follow Shelley on that point, for instance Rachel Ankeny has written that model systems "usually encompasses not only the organism but also the techniques and experimental methodologies surrounding the organism itself" (Ankeny 2007, p.47).

strategy, which the title stresses: the paper attempts to model the effect of maternal infection on the developing human central nervous system, by means of a radically different system. The most obvious difference is the use of a developing rat instead of a developing human. There is, however, another important element of modeling: the researchers did not use real infectious agents (which can have variability in their effect and require containment measures), but instead relied on Lipopolysaccharide (LPS). LPS is a molecule derived from the outer membrane of some bacteria, and it is widely used in research because of its potent capacity to activate immune response. In effect, LPS produces an immune reaction akin to that of an infection even though, strictly speaking, there is no infection. Their modeling strategies therefore rely on a double analogy: between rat and humans, as well as between LPS and infection. As in Porsolt's test, there should be no doubt that the LPS is also modeling.

The same can be said of most, if not all animal models, although it is not always as obvious. When scientists test, for instance, whether sustained ingestion of a substance causes cancer, the experimental setting is very similar to the target setting: in both cases, the individuals (whether mice or men) eat the substance, and in both cases they develop cancer at a higher incidence (or not). In fact, however, the situation is not essentially different from Porsolt's test. To give a simple example, mice obviously do not eat as much as humans do, and the amount eaten by the mouse is assumed to model either the amount that a human might eat or the metabolic concentration it may lead to.

This being said, Shelley's account is too inclusive, for it considers also the drug being tested as part of the model. In his discussion of Robert H.H. Koch's famous experiments on tuberculosis, Shelley writes:

> "Koch did not re-use the same guinea pig model of tuberculosis. Instead, he modified an old model for the testing of tuberculin by changing the testing conditions, especially the 'tuberculin' itself." (Shelley 2010, p.297)

Shelley's position leads to a major problem: by conflating models with experiments, it prevents models from being evaluated beyond the exact experiments in which they have been used. Instead, I would argue that the nature (rather than, say, the dose) of tuberculin itself is not part of the model on the ground that it is not standing for the drug he would give to patients: it is, instead, the very same thing. And to make model individuation more

useful, I believe it is important to restrict it even further.

1.3.1 Model individuation

An animal model, therefore, is a system in which the animal is but one component, which does not necessarily imply that all components are equal. The relationship between a model and its target implies a mapping between the components of both systems. However, not everything scientists use is part of the model, for not everything corresponds to something in the target system. The fact that the mouse lives in a cage, for instance, has no human counterpart – it is not in itself modeling anything, but it is simply a practical necessity. Inasmuch as components of the modeling system *stand for* analogous components in the target system, they can be said to model in their own right. This does not necessarily imply that they are not part of a single model. As Giere and Weisberg point out, modeling implies a mapping between different components of both systems, but how are we to say whether the rat and the LPS represent two models, or two components of the same model?

I would suggest that *two elements are part of the same model insofar as their mapping with elements of the target system are interdependent*. This will obviously leave grey areas, but for the moment it should be enough to work with. Consider again Porsolt's forced swim test: the tube filled with water, the animal and the procedure are all part of the model because none of these elements, on its own, has anything to do with depression. Mice (most probably) do not have suicidal thoughts, and in any case the mouse is not by itself a model of depression, but becomes one coupled with Porsolt's setup. The mapping is not even robust to changes in other elements: Porsolt's setup would be totally irrelevant with other animals such as fish or humans, who would both not even try to escape in the first place.

On the other hand, when one dissects a mouse to describe the way blood is irrigating tissues, the fact that the mouse models humans blood irrigation does not depend very much on the tools used for the dissection[18]. Whether a given system, or a component of

[18] There is obviously no denying that some dissection tools are better than others; instead the point is that the adequacy of these tools and the adequacy of the mouse are not strongly dependent on each other. Consider, for instance, the fact that the computed tomography scanner used for mice is considerably smaller than the one used for humans: while one could say that one stands for the other, the modeling capacity of the mouse as a model is independent of which scanner one uses.

a system, ought to be called a model by itself depends on whether it can be evaluated as a model on its own, i.e. whether its mapping with the target system is robust to changes in elements outside of the system.

Following Shelley's point, the common practice of referring to an organism as a 'model' seems mistaken. It would merely be an unfortunate conflation of the meaning of the term model with its coincidental extension: because the laboratory mouse is associated with so many models, it came to inherit a label that does not refer to it. However, the tentative criterion proposed here implies that whether an organism can be said to be a model on its own will depend on the context.

An important aspect of contemporary biomedical research is that organisms increasingly *contain* the experimental setting. Consider the example of the OncoMouse®, the first commercially available, genetically-engineered mouse strain spontaneously developing breast cancer within a few months of birth[19]. Before genetic engineering, cancers were generally induced in animals through exposure to carcinogens or irradiation. In this context, exposure to the carcinogenic agent is modeling human exposure to (very roughly) similar agents. In the OncoMouse®, the kind of genetic lesion at the origin of many cancers is already included in the mouse's genome[20]. As such, the mouse is not only standing for a human being, but it is also standing for the events of DNA damage or copy errors that cause cancer in humans. This does not only turn the organism into a technology (Gachelin 2006), but gears it towards some applications:

> "The intact organism itself is turned into a laboratory. It is no longer the extracellular representation of intracellular processes, i.e., the 'understanding' of life that matters, but rather the intracellular representation of an extracellular project, i.e., the deliberate 'rewriting' of life." (Rheinberger 2000, p.25)

The phenomenon is not limited to genetic engineering. Already in the early 20th century, model organisms where being intensely standardized (Rader 2004). Genetic standardization through inbreeding literally turned *Mus musculus* into a tool for the study of the genetic

[19] The OncoMouse®, developed by Philip Leder and Timothy A. Stewart using tissue-specific expression of the RAS oncogene, was commercialized by Du Pont starting in 1988. The OncoMouse® was also the first genetically modified animal to be patented (patent EP0169672A1, filed in 1985 and held by The President And Fellows Of Harvard College). The application has led to extremely interesting debates, especially in Europe and in Canada – see especially the 2002 judgement of the Supreme Court on the notion of 'composition of matter', in *Harvard College vs. Canada (Commissioner of Patents)*.

[20] This being said, it was pointed out to me that RAS mutations were actually very rare in breast cancers. As such, the OncoMouse models carcinogenic lesions rather than, as it purports to, breast cancer lesions.

contributions to organismal phenomena. As such, it materialized a whole epistemic culture within the organism, making it an intrinsic *model for* (Keller 2000).

Nevertheless, it is most often misleading to consider the animal as being transplanted alone across research contexts. This is particularly true of 'model organisms'. In what is probably the most widely accepted account of model organisms, Rachel Ankeny and Sabina Leonelli (2011) distill two features distinguishing model organisms from other organisms used in research. One of them is that model organism-based research aims at integrating a wide variety of approaches, phenomena and technologies, in order to understand "whole, intact organisms, in other words for a range of systems and processes which occur in living organisms, including genetics, development, physiology, evolution, and ecology." (Ankeny and Leonelli 2011, p.318). Such goals are not properties of the organism *per se*, but of the research community. Rather, a model organism is a broader ensemble encompassing technologies[21], communities, practices, and organisms. This is obvious in the importance that institutions such as the Jackson Laboratory or the Mouse Genome Institute (MGI) had in the constitution – and current functioning – of *Mus musculus* as a model organism. Model organisms, therefore, are not organisms that have the additional property of being models, but models which have the organism as its central component.

1.4 Models or experimental systems?

As we saw in section 1.1.1, models have the characteristic function of mediating between theoretical representation and reality (data or phenomena). Although biomedical models are themselves a reality, they are an idealized reality. They can therefore be said to mediate access to some 'natural' reality, in the sense of facilitating access to it – not only intellectual/perceptual access (as in Bailer-Jones 2002), but also material/empirical access. This, however, is not limited to models, and is equally true of experimental systems broadly conceived[22]. Biomedical models offer stabilized phenomena, and thereby provide "a stable

[21] It is interesting that even for tasks such as genetic engineering, which *a priori* seem most common across organisms, many model organisms have their own peculiar technologies: morpholinos, for instance, are used only in some organisms, and different systems for conditional expression are more common in some model organisms than others.

[22] "The experimental conditions 'contain' the scientific objects in the double sense of this expression: they embed them, and through that very embracement, they restrict and constrain them." (Rheinberger 1997, p.29)

target of explanation" (Keller 2002, p.115). They also provide material access to it in the sense that this superimposed nature is generally more easily manipulated and, as we will see through the next chapters, it brings specific dimensions of nature into visibility.

Construing biomedical models in this way suggests the existence of an original, natural reality which is being mediated. However, "nature itself only becomes real, in a scientific and technical sense, as a model" (Rheinberger 1997, p.108). This is not to deny that there is a reality independent of science: what is denied is that this reality is organized into phenomena. Phenomena – both modeled and being modeled – are constructed (Hacking 1983), and for this reason "Neither models nor reals are givens" (Rheinberger 1997, p.91). This has already been hinted at in section 1.2.2, and will be explored in Chapter 4. For the purpose of the present discussion, what this means is that biological reality is always mediated, and it is for this reason that the distinction between biomedical models and non-models has been so elusive. For all the characteristics we have found to be possessed by biomedical models turn out to be more broadly possessed by experimental systems: they mediate between theory (or, at any rate, theoretical constructions) and reality, and do so because elements of the experimental system *stand for* elements of the phenomena (e.g. LPS administration to *in vitro* B cells stands for an infection).

Experimental systems are (following an account by Rheinberger 1997) relatively stable, organized systems composed of materials (which may include organisms), instruments, procedures (which may include conceptual procedures), geared toward answering "questions that the experimenters themselves are not yet clearly able to ask" (Rheinberger 1997, p.28). This means that even from the point of view of model individuation, the view on models applicable to biomedical models coincides with experimental systems. In the previous section, I have argued that whether a given system ought to be called a model on its own depends on the extent to which its modeling capacity is autonomous. Likewise, Rheinberger describes experimental systems as "the smallest integral working units of research." (Rheinberger 1997, p.28) Of course, no experimental system works on its own, and in fact Rheinberger suggests to define them ecologically (Rheinberger 1997, p.135). But they are to be epistemological *units* of research precisely because they are islands of robustness in this constantly rearranged system of representation. Robust not in the sense that they are

predictable – they must not be, at least not entirely, in order to be productive – but in the sense that they are reproducible, which means repeatable with difference or in different contexts. Note that this does not exclude the possibility of there being overlapping experimental systems, just like organs and cell types represent two overlapping and yet robust units of biological organization.

Thus, construing the notion of model in such a way as to apply to biomedical models leads to its collapse with the notion of experimental system. As a consequence, either we should drop the notion of model in this context, or at least stop considering models as opposed to some non-modeled referent (i.e. direct experimentation). Furthermore, as I have argued in section 1.1.3, the adequacy of a representation is not reducible to a relation between the represented and that doing the representation, because what the representation actually does is to situate the represented on a sort of map – on a broader system constituted by other relations of representation. These are two important departures from most of the literature on biomedical models, and the following chapter will reiterate these claims and their consequences in more concrete and more specific contexts of biomedical research (chapter 2). The other chapters will then be concerned with the articulation of an alternative framework for the use of biomedical models. In the context of biomedical models, we should consider the clinical setting not as a target in a unidirectional extrapolation, but as another model system (or experimental system) within a complex network of what I will call distributed modeling. Considering this network allows for a proper understanding of the precise functioning of specific subsystems within the research activity, and of their mutual dependencies.

Chapter 2

Evaluating biomedical models

2.1 Introduction

The goal of this chapter is to point out some major problems in the way biomedical models have generally been conceptualized. The chapter (section 2.2) begins by discussing the scope and limits of an influential body of literature criticizing the value of animal experimentation. Although the evidence raised suggests that animals are very poor biomedical models, I will argue that the critique fails to pay adequate attention to the functions played by biomedical models, and presupposes a narrow predictive usage of biomedical models which seldom obtains. The second part (section 2.3) presents a concrete example of what is generally considered a paradigmatic example of the predictive usage of biomedical models, and shows that even in such an example the biomedical models are not used in this narrow way. Finally, in the last part (section 2.4) I build on this discussion to lay out three problematic assumptions that have hampered most philosophical accounts of biomedical models.

The first part of the chapter is an assessment of the main epistemological arguments against animal testing (section 2.2.1), concentrating on the work of Niall Shanks and C. Ray Greek (especially Shanks and Greek 2009). These criticisms have a very limited scope, applying to a very narrow usage of biomedical models which they label the 'predictive use' of models. However, before showing the inadequacy of this view, I will argue that even within this narrow scope, their argument can only be conclusive with respect to specific pragmatic goals and alternatives towards that goal. I rely on some basic notions of test

statistics to describe some problems related to the evaluation of biomedical models (these notions are explained in appendix A, where I also criticize a misleading argument that is commonly held by critics of animal models). I then discuss how different notions, especially the notions of sensitivity and specificity, are relevant at different degrees depending both on the base rate and on the broader aims of a screen (section 2.2.4). A first lesson from this overview is that in general models do not work in isolation, but in conjunction, which has important implications for their evaluation.

Beyond these issues, I will argue that Shanks and Greek's view on biomedical modeling is largely inadequate. In order to show this, I rely on a concrete example expected to fit squarely into the 'predictive use' of biomedical models Shanks and Greek discussed (section 2.3). I discuss what is probably the paradigmatic case of a drug discovery screen: the massive screening for chemotherapeutics conducted since 1955 at the US Cancer Chemotherapy National Service Center (CCNSC). In reviewing some of the important historical developments of the CCNSC, I highlight a major shift in its conception and use of models (sections 2.3.1 and 2.3.2). In the 1980's, the screen was reconceptualized from a prediction/exclusion tool to a more complex form of inquiry, although in practice it had functioned in this way even earlier. This shift represents a radical departure from the narrow predictive use described by Shanks and Greek.

Finally, building on points raised in the first two sections of the chapter, I lay out three major assumptions which have hampered most philosophical accounts of biomedical models (section 2.4) :

1. that biomedical models function as surrogates for human beings (see section 2.4.1);
2. that extrapolation runs undirectionally from model to target system (see section 2.4.2); and
3. that modeling and extrapolation can be understood solely by looking at the model-target dyad (see section 2.4.3).

The next chapters will then attempt to solve these shortcomings. Chapter 3 will focus on the first assumption, while the following chapters will discuss the others in more detail.

2.2 Screens and predictivity

2.2.1 The critique of animal testing

The use of animal testing in science has undergone severe criticism in the last few decades, partly because of growing ethical concerns (see section 0.3.1), but also for mere reasons of efficiency, given the diminishing returns of investments in biomedical research (see section 0.3.2). Some critics have focused on empirical arguments showing the inefficiency of animal experimentation (see especially Knight 2008, 2011), while others have appealed to both empirical and theoretical arguments (LaFollette and Shanks 1995, 1996, Shanks and Greek 2009, Shanks et al. 2009, Greek and Shanks 2011). As there is a lot of overlap between these authors, I will not discuss them all, but will instead sketch an overview of the main arguments, starting with the theoretical arguments.

LaFollette and Shanks (1995, 1996) introduce a distinction between two uses of animal models: as Hypothetical Analogical Models (HAMs), whose results provide food for thought or hypotheses to be tested on the target system, and Causal Analogical Models (CAMs), whose results are predictive – i.e. they can be directly extrapolated to the target system. Their basic argument is that scientists use animal models as CAMs, but that the models do not have the requirements for this usage. LaFollette and Shanks (1995) summarize notion of CAM in the following way:

> "To put it more formally, CAMs fit the following schema of all analogical arguments: *X (the model) is similar to Y (the subject being modelled) with respect to properties [a, ..., e]. X has additional property f: Wile f has not yet been observed directly in Y, it is likely that Y also has the property f.* Since CAMs are a subspecies of analogical arguments in which (some of) the premises and conclusions involve causal analogical claims, the CAMS must satisfy two further conditions especially relevant to its causal dimensions: *(1) the common properties [a, ..., e] must be causal properties, which (2) are causally connected with the property f we wish to project – specifically, f should stand as the cause(s) or effect(s) of the features [a, ..., e] in the model.*" (LaFollette and Shanks 1995, p.147, original emphasis)

On this account, in order for an extrapolation to be warranted, "there must be no causally relevant disanalogies between the model and the thing modelled" (LaFollette and Shanks 1995, p.147). As there are most often (if not always) causally relevant disanalogies between humans and animal models, they conclude that scientists' use of animal models as CAMs

is unwarranted. Shanks and Greek (2009) provide an impressive list of relevant differences between humans and higher mammals, along with theoretical reasons to expect such differences (mostly evolutionary and developmental reasons). Furthermore, they have argued that "the causes and effects of the events that a complex system experiences are not proportional to each other" (Greek and Shanks 2011, p.542), so that even small differences can have a considerable impact at the level of the organism.

I have no quarrel with the fact that animals are not humans. The structure of the argument is plausible, and I will not raise any objection against it here. It is however important to highlight its scope. First, as many authors have noted, the argument applies equally well to extrapolation between humans:

> "Not only are relevant differences across species inevitable, but dissimilarities are also extremely common *within species* and even for a *single organism* at different stages of its life. [...] This, if the strict criterion of CAM-hood proposed by LaFollette and Shanks were accepted, not only would extrapolation from animal to human be illegitimate, but so would extrapolation from humans to other humans." (Steel 2008, p.93)[1]

It is puzzling that Shanks and Greek (2009) spend tens of pages discussing human variation and adverse effects, without addressing this crucial problem: if the fact that animals are not CAMs prevents them from being used as predictive, why should clinical trials and other clinical studies be taken as predictive?

As I will argue in Chapter 4 (section 4.2.1), the ground for animal experimentation is tied to the very possibility of experimental medicine. For the moment, however, I wish to discuss another aspect of the scope of the argument by Shanks and Greek. The argument targets a very particular kind of animal modeling:

> "There is no doubt that careful biological studies of rats and mice can help clarify the general contours of mammalian biology. Such studies can also play a valuable heuristic role by prompting new ways of thinking about human biological problems of interest. The issue we are concerned with is this: notwithstanding these cautions, are animal models predictive of human outcomes in, say, toxicology, drug discovery, and the study of the causes and cures of human diseases?" (Shanks and Greek 2009, p.29)

The criticism does not therefore pertain to the use of animals in research in general, but to what the authors call 'predictive animal modeling' – namely the assumption that a result

[1] See also the discussion about within-species variation in Chapter 1, section 1.2.1.

obtained in animal studies is conclusive for humans, rather than merely leading to (at best likely) hypotheses. Shanks and Greek (2009) argue that

> "This is how animal models are, in fact, used in the context of drug testing and studying human disease. Animals in the case of predictive models are clearly used *as substitutes for human subjects*." (Shanks and Greek 2009, p.117, emphasis added)

In other words, the authors assume that "in the context of drug testing and studying human disease", animal models are used as *surrogates* for human patients, and ought to respond as humans would have. It is with this presupposed characterization of the way animals are used in science that I take issue. I will argue that most of animal modeling does not fit the picture of 'predictive animal modeling', and that animal models need not be surrogates for human patients.

Shanks and Greek (2009) repeatedly emphasize that they are not concerned with basic research. Relying on the rather naive definition of the Organisation for Economic Cooperation and Development, they adopt the following account of the division between basic and applied research:

> "Basic research is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundation of phenomena and observable facts, without any particular application or use in view. By contrast, applied research consists also of investigations undertaken to acquire new knowledge, but it differs from basic research by being directed primarily to the achievement of particular aims and objectives." (Shanks and Greek 2009, p.370)

As Kitcher (2003) and others have shown, there is no 'pure' science; furthermore, even within this continuum the distinction between basic and applied research has become increasingly blurry in biomedical research (see for instance Cambrosio and Keating 2003). Almost by definition, biomedical research is applied, for it ultimately has medical aims. These aims can be more or less concrete, but because of funding requirements even very fundamental projects generally get framed as applied, and the scientists themselves often oscillate between the two, jumping on opportunities to turn one into the other. For these reasons, an artificial division can be misleading.

Shanks and Greek (2009) are sometimes more precise, writing that they are concerned with the predictive usages of animal models such as toxicology testing or drug discovery.

This greatly reduces the scope of their argument, for fundamental biological studies account for nearly 40% of animals used for scientific purposes in the European Union, in contrast to mere 8.7% for 'Toxicology and other safety evaluation' (European Comission 2010, p.7). Notwithstanding these issues, my claim is that even these paradigmatic examples of applied research need not assume neither surrogacy nor conclusive predictivity. My main example to this effect will be (in the second part of this chapter) the massive drug screening of the Cancer Chemotherapy National Service Center (CCNSC).

Before turning to this example, I would like in the first part of this paper to assume for a moment that the characterization of biomedical models as surrogates is correct, in order to show that even in this context, the requirement for conclusive predictivity is misleading.

Despite regular slips of language from the part of scientists (of which Shanks and Greek 2009 have made quite an inventory), the claim that animal studies are not completely predictive of human outcomes is nowadays uncontroversial and well acknowledged by scientists. Note that this was not always the case: for instance, Robert Koch (1843-1910), recognized as the founder of modern bacteriology, at first thought it unquestionable that his medicine working in Guinea pig would also work in men, and considered the translation to be merely a routine application of an already developed intervention; in is only after the failure in humans that his retrospective interpretation changed[2].

Contemporary scientists are well aware that animal models are not conclusively predictive of human outcomes, but nevertheless rely on those predictions. An important problem is that critics of animal experimentation often write as if animal testing was being used *instead* of human experimentation. For instance:

> "The most direct way to test hypotheses about humans is to conduct tests on humans. [...] However, although there are established retrospective and prospective research methodologies for such studies, many biomedical researchers advocate using non-human models as CAMs *instead*." (LaFollette and Shanks 1995, p.144, emphasis added)

Animal testing, however, is seldom meant to *replace* human testing: instead, it is first and foremost a tool for the allocation of resources.

[2] Roelcke for instance notes: "Das, was sich in der zweiten November-hälfte 1890 als Routineverfahren darstellte, wurde zwei Monate später als therapeutisches Experiment am Menschen eingeordnet." Roelcke (2009, p.27) Other scientists, such as Claude Bernard, also believed in strict predictivity, although of a more complex kind, and I will return to his approach in Chapter 4.

2.2.2 Multi-step screening

There are several ways to assess the quality of a test or screening procedure, and some of them are more relevant than others depending on the broader goal of the procedure. The basic statistical concepts are summarized in appendix A (section A.1). Shanks and Greek are well aware of these notions, and make the more specific points that "animals (when taken as a whole) are very sensitive but not very specific" (Shanks and Greek 2009, p.291; see also Olson et al. 2000 who reach a similar conclusion), and that each animal on its own lacks sensitivity. Shanks et al. (2009) judge these performances against strict standards generally applied to medical devices:

> "in order for a test to be useful given the demanding standards of medical practice, in this case tell us if the patient actually has liver disease, it needs to be have PPV and NPV in at least the .95 to 1.0 range." (Shanks et al. 2009, p.5)

Such standards, however, are necessarily tied to what we do with the test. Furthermore, as Shelley (2010) pointed out, the validity of animal studies should not be evaluated against a fixed threshold, but against the alternatives we have. Perhaps, the thought goes, animal studies are simply the best we have (or can afford). To this general line of argument, Shanks and colleagues have answered that

> "Astrology is not predictive for foretelling the future therefore we criticize such use even though we have no notion of how to go about inventing such a future-telling device." (Shanks et al. 2009, p.18)

But this is mistaken, for we do have a future-telling, or in fact an all-purpose device: coin-tossing. The whole point of criticizing astrology is that it does not do better than chance; would it, we would acknowledge it. Anything that does better than chance is potentially useful, and as I will argue in a moment we have reasons to believe that animal studies do better than coin-tossing.

Animal testing ought to be judged in comparison with its alternatives, but this is not to say, however, that animal studies ought to be discarded if there is a better alternative. The reason is that for most practical purposes, a test does not stand on its own, and this is especially obvious in the context of mass screening. Consider for instance the famous 'Pap-smear' (or Pap-Test), named after its inventor Georgios Papanikolaou and used to

screen for cervical cancer. Despite a total lack of interest by the scientific community, Papanikolaou laboriously studied countless vaginal smears for decades in an attempt to describe the menstrual cycle. At that time, his endeavour was descriptive, but when it was discovered that early cervical cancers presented anomalies on the Pap-smear, it ceased to be an object of inquiry (an epistemic thing) and started becoming a technical tool for the detection of cervical cancer. Importantly, an anomaly detected on the Pap-smear does not immediately lead to intervention. Instead, patients showing abnormalities undergo a biopsy. The biopsy and following study by a pathologist is both more sensitive and more specific than the Pap smear, but it is longer, more costly and more complex to perform, preventing its regular and widespread application to the whole population. In this context, the job expected from the Pap smear is to reduce the number of unnecessary biopsies. As a consequence, assuming a high sensitivity, even a very low PPV – say 0.2 – would be highly cost-effective: since about 5% of Pap Smears show some abnormality (less than 1% are indicative of cancer or high grade lesions), this would still mean to reduce the number of unnecessary biopsies at least by a factor of 1/20 (as a matter of fact, the Pap smear is quite specific – see Coste et al. (2003)). From an epistemic point of view, it would be ridiculous to claim that the Pap Smear should be abandoned on the ground that a better standard exists, because the two together achieve roughly the same effect at considerably lower costs[3].

When Shanks et al. (2009) write that results from animal studies are only hypotheses of human outcomes, and that "[t]he prediction that the hypothesis entails must then be tested" (Shanks et al. 2009, p.2), they can mean two things. They can mean that we do not have absolute certainty as to whether the result can be extrapolated, which I believe is uncontroversial. But they can also mean something stronger, namely that the results of animal studies must systematically be tested in humans in order to be of any use in fields dealing with prediction, such as is the case in toxicology and drug discovery. This would amount to saying that anything short of being the gold standard should be dismissed, which would be nonsense. Once again, the Pap Smear is not entirely predictive of neoplasia. Nevertheless, confirmation of the result is not sought for negative results. It would be

[3] Depending on the sensitivity of the test, there might however be ethical issues in relying on the lower standard, but often such a multi-step screening is the only feasible possibility.

impossible to regularly test every woman through the more accurate methods, and it is for this reason that the Pap Smear is a success story.

The situation is exactly the same in the case of drug screening, in which pre-clinical studies are but a mean of reducing the amount of unsuccessful clinical studies. Consider the following passage quoted by Shanks et al. (2009):

> "Currently, nine out of ten experimental drugs fail in clinical studies because we cannot accurately predict how they will behave in people based on laboratory and animal studies" (U.S. Secretary of Health and Human Services in 2007, quoted in Shanks et al. 2009, p.4)

As bad as this 9/10 sounds, an argument solely based on the attrition rate[4] cannot be used to assess the utility of a test procedure. The danger in interpreting such a claim is to ignore how rare effective drugs actually are, an error known as the base-rate fallacy (see appendix A, section A.2). Assuming that 1% of the compounds tested do in fact have the desired effect in animal studies, this means that 1000 tests were performed in mice, and only 10 in humans, to find the drug. Unless sensitivity is *very* poor, this means that we will have saved hundreds of failed clinical trials. Within the bigger picture, 1 true positive out of ten might be a very desirable thing.

2.2.3 The missing base rate

Calculating both specificity and sensitivity requires knowledge of which (or at least how many) compounds are in reality effective or ineffective. In general, this is approximated using a gold standard: for instance, the specificity and sensitivity of the Pap Smear in screening for cervical cancer is evaluated by comparing its results with a pathologist's deeper analysis of a biopsy. As biopsies are not prohibitively costly or dangerous, it was possible to run extensive trials comparing the two. A major problem in drug discovery and toxicology is that this is most often not possible.

Because phase III clinical trials are considered the gold standard, a drug that passes this

[4] In drug development, the attrition rate is the rate at which the number of candidate compounds shrink through the development process. Attrition rate can refer to the passage between any steps of the development progress, although typically it will be the proportion of compounds that make it to the market, relative to the proportion of compounds entering clinical (or preclinical) studies. The high attrition rate, especially in cancer drug development, is perceived as a major problem in the industry (Arrowsmith 2012, Scannell et al. 2012).

final test is considered a true positive, and can be used retrospectively to assess the earlier steps. The problem, however, is that only the compounds that have been retained until this final step are known to be true positives or false positives, so that we do not know, among all the rejected compounds, how many were false negatives. In other words, in general we cannot assess the proportion of effective materials at any step before the final one. Unless one knows what proportion of all tested compounds are positive, it is impossible to infer the sensitivity, specificity, or the accuracy, of the screen or of any of its steps. Because in general only positive hits are further studied, the only thing we can assess is the Positive Predictive Value (PPV)[5] of the pre-clinical screening as a whole.

The problem of the missing base rate was raised in the context of the CCNSC's drug screening, which I will discuss in section 2.3. To solve this issue, the famous clinical pharmacologist Luis Lasagna[6] suggested to test on humans some of the compounds that did not pass the pre-clinical screen:

> "I see no way, for example, to decide about 'true' or 'false' negatives without testing on man at least some of the compounds that seem worthless for treating animals. One may be unwilling to do this, or it may seem a waste of time, but one should not make judgments about what is or is not being discarded in a screen without at least sampling the discards to see whether the screen's rejections are altogether reliable." (Lasagna 1958, p.939)

This was not done, most certainly because of ethical concerns. The general scientific opinion instead favored a poorer but arguably ethically more acceptable alternative, namely to ask how well the screen would 'discover' drugs which we already knew to be active. This study came to be known as 'the Gellhorn-Hirschberg report' (Gellhorn and Hirschberg 1955). Note that there were substantial reasons to resist this alternative: to start with, the cancer drugs we knew to be active were far from a random sampling of effective compounds – in fact they all belonged roughly to a single category of chemotherapy agents. On top of being statistically inadequate, this meant that calibrating the screen on this set would lead the screen to discover more of the same. As Lasagna notes,

> "A second problem is whether a screen that actually can select the drugs we already have is desirable. One could take the position, for example, that the

[5] Terms related to test statistics are defined in appendix A, section A.1.

[6] Lasagna is generally considered the founder of clinical pharmacology. He made major contributions to clinical trial methodology and placebo response research, and advocated stricter regulations for the approval of drugs.

> best drugs we have are far from adequate, since they are highly toxic compounds that only occasionally show greater toxicity for tumor tissue than for normal tissue; consequently, a screen oriented around such agents may merely turn up additional compounds of the same unsatisfactory type." (Lasagna 1958, p.940)

Nevertheless, for lack of a more acceptable alternative the results of the Gellhorn-Hirschberg report were used to select the model systems for the CCNSC screen (the results were rather poor, but the models most predictive of known drugs were selected).

Similar studies were used by critics of animal testing to evaluate the sensitivity of animal models to human carcinogens:

> "If we consider long term feeding or inhalation studies and examine all 26, only 12 (46.2%) have been shown to cause cancer in rats or mice after chronic exposure by feeding or inhalation. Thus the lifetime feeding study in mice and rats appears to have less than a 50% probability of finding known human carcinogens. On the basis of probability theory, we would have been better off to toss a coin." (Salsburg 1983, cited in Shanks and Greek 2009, p.270)

Such studies generally suffer from a very small sample size and are unrepresentative of the tested population. Furthermore, they cannot allow us to assess specificity, because they contain no information about the proportion between true negatives and false negatives; in order to assess it, it is necessary to combine them with attritions rates obtained from very different samples (and sampling methods). If animal studies were truly worse than coin-tossing in both respects, this would impose a radical rethinking of pre-clinical research. But the critics have failed to convincingly show that this is case, and some comparisons with coin-tossing have explicitly committed a base-rate fallacy (see appendix A, section A.2). This being said, there is no denying that the overall performance of animal studies for predictive purposes is very poor (for a careful evaluation of the empirical evidence, see Knight 2008, 2011), and there may be specific fields in which animal studies are systematically misleading.

In any case, it is important to point out that whether or not we are justified in using animals for predictive purposes is not strictly a question of statistics. It is a question of balancing costs and benefits (Knight 2011): economical costs (such as in the case of the Pap-smear), but ultimately also the costs of animal lives with respects to the risks we, as prospective patients and consumers, are ready to take.

2.2.4 Sensitivity and specificity

Very often in practice there is a trade-off between sensitivity and specificity, so that making a test more stringent to increase specificity often leads to decreases in sensitivity. While there is no *a priori* reason to favor one over the other, in most cases there is a rationale for doing so based on the practical consequences of the test. To give a simple example, if we have a test procedure to know whether someone has a very dangerous infection, we want to maximize sensitivity even at the extent of some specificity. The reason is that while widespread prescription of antibiotics is extremely problematic, the costs (both moral and economic) of failing to prescribe it to one individual who needs it far outweigh the costs of giving antibiotics to one individual who does not need it. Likewise, the presumption of innocence in philosophy of law is based on the idea that it is better to let a criminal run free than to condemn an innocent.

It is interesting that a similar reasoning is also at play in research methodology. In applying a Neyman-Pearson testing procedure, one has to choose significance thresholds – or to chose the acceptable risks of committing type I and type II errors. Interestingly, these are not valued equally: typically, the risk of committing a type I error (rejecting the null hypothesis when it is in fact true) is set to 0.05 or 0.01, while the accepted risk of committing a type II error (accepting the null hypothesis when it is in fact false) is often as high as 0.2. In other words, scientists find it more problematic to wrongly accept a result as significant than to wrongly dismiss a result as non-significant. Part of the explanation for this is sociological: results deemed significant are published, and as such are often the only ones the scientific community will be aware of, and it is damaging for a scientist's reputation if his results cannot be replicated.

In the context of a toxicity screen, I would argue that it is advisable to favor sensitivity (to toxicity) over specificity, because in most circumstances, the consequences of allowing a toxic compound on the market outweigh the consequences of banning a safe compound. The reason is not so much because toxic effects are necessarily more important than beneficial effects of a compound: it might be that the costs of failing to release a very good drug to the public are worse than those of letting some toxic compounds on the market. Instead, the reason to favor sensitivity depends on the asymmetry of the system in which the test

is embedded: those who have the most interest in banning a dangerous compound – the consumers – have very little power to detect them, and those who have most power to detect them – the industry – have limited interest to do so. If mandatory toxicity screens declare a compound non-toxic, the industry will not have much interest in further studying its potential toxicity. However, if a compound or a drug is so highly promising that its banning would do much damage to society, the industry will have an interest to go beyond the mandatory toxicity screen. They will study the mechanisms and dynamics involved, and if there are good reasons to believe that, despite the toxicity screen, the compound should not be toxic in humans, they will build a good case for it. Once more, a test does not stand on its own.

The situation is slightly different in the context of drug discovery, and to see this I will now turn to a concrete example.

2.3 The Cancer Chemotherapy National Service Center

In the first part of this chapter, I have argued that the critique of Shanks and Greek (2009) applies to a very limited subset of the animals used in biomedical research: those used as surrogates for human patients, predicting human outcomes from animal outcomes. I have argued that even this so-called 'predictive usage' does not require conclusive predictivity, and that the value of a test cannot be assessed on its own.

In this second part of this chapter, I will further restrict the scope of the arguments of Shanks and Greek by showing that even the paradigmatic examples of drug discovery do not rely on such a predictive usage. I will be concerned with the high-throughput drug screening performed at the Cancer Chemotherapy National Service Center (CCNSC) since 1955 (the major developments of which are discussed in section 2.3.2).

2.3.1 Screening in the context of drug discovery

As I argued in section 2.2.2, screening is not a practice of strict prediction, and as it was understood at the beginning of the CCNSC, it was a practice of gradual exclusion. Drugs giving positive results in the animal are not assumed to be efficacious for human

treatment, but are studied further, up to clinical trials. In this context, the job of screening programs is to reduce the number and proportion of uninteresting compounds. As the medical statisticians Peter Armitage[7] and Marvin A. Schneiderman wrote in a discussion surrounding the screen:

> "By 'mass screening program' we mean a procedure for examining a large series of materials in such a way that the minority are retained for further investigation while the majority are rejected or set aside." (Armitage and Schneiderman 1958, p.896)

This is clearly a question of material resources: clinical trials are obviously better at finding drugs than pre-clinical studies, but they are also much more costly, both morally and economically. In general, the same is true of the different phases of clinical studies. Consider, for example, the following passage on anticancer drug development:

> "The typical costs of conducting phase I, II, and III clinical trials are US$200,000, $1 million, and $10 million, respectively. Usually, a novel anticancer agent is tested in one to four phase I trials to evaluate distinct schedules. Phase II trials will be conducted in five to ten tumor types. This step from phase I to phase II trials requires a 10-fold increment in investment. The limited public and private funding for clinical investigations mandates that agents be stringently selected for phase II development." (Thambi and Sausville 2004, p.364)

Note the symmetrical relationship between the estimated costs of each step, and the number of such tests performed. In a similar way, in pre-clinical screening cheaper tests are generally used earlier in the screening process. Corbett et al. (2004) notes that a lesson learned from the early NCI screens is that "a supersensitive model will divert and consume resources at an alarming rate" (Corbett et al. 2004, p.100). Screening is a search strategy, and multi-stage screening is a method to maximize cost-efficiency. Indeed, Armitage and Schneiderman (1958) mention that mathematical models were made to estimate the optimal number of steps (the models suggested a 3-stage procedure).

The common metaphor given is the one of progressively narrower sieves which gradually eliminate compounds until only viable hits are left. While the metaphor gives a rough

[7] Although he has had a much broader career in the United Kingdom, Armitage is mostly known in the field of cancer research for the famous Armitage-Doll multistage model of carcinogenesis (Armitage and Doll 1954, 1957), the basic logic of which still underpins today's understanding of carcinogenesis. Schneiderman instead had a career of over 30 years at the NCI, and contributed most importantly to the epidemiological link between cancer and hazards such as smoking or sun exposure.

idea, actual screening is more efficient, because the tested compounds are not entirely independent and therefore not entirely random. In a nutshell, mildly positive leads suggest to test related compounds[8]. Notwithstanding these additional complexities, the value of a given step of the screen is generally estimated through the extent to which it increases the concentration of true positives in the test population:

> "Of the screen as a whole, we should require not only that the proportion of effective materials be increased progressively with the several stages of screening, but also that it be increased to a reasonably high value. An increase of this proportion from 1 in 100,000 of entering materials, for instance, to 1 in 50 after several stages, would indicate that still more screening was necessary." (Armitage and Schneiderman 1958, p.896)

In other words, the requirement is that the Positive Predictive Value (PPV) increases at each step of the screen. This is most likely not an ideal requirement, but a practical requirement: as we saw earlier, this is something that can be estimated without knowledge of the base rate, on the basis of the attrition rate at the clinical level. A drug that passes clinical trials is considered a true positive, and used retrospectively to assess the pre-clinical screen. Still, this presents the difficulty that it can only be used to judge the pre-clinical screening process as a whole, for compounds that are rejected between the first and the second step of pre-clinical screening will almost never be tested clinically.

The scientists of the CCNSC relied on the information they had at their disposal. Assuming, however, that good knowledge of the sensitivity and specificity of a screen (or screening step) had been available, would it be reasonable to favor either of sensitivity and specificity? As stated above, the rationale for doing pre-clinical study is to reduce the number of unsuccessful clinical trials, which is equivalent to increasing the PPV. This suggests that we should not care about False negatives, but in reality whether we should depends on the circumstances. It might be useless to have a screen with a very high PPV, if this screen gives only one candidate every decade – even if the candidate is a highly

[8] "For example, let us propose an inventory of 300,000 agents (a medium-size inventory for a pharmaceutical company), and propose that three novel solid tumor-active clinical candidates exist in that inventory. If we test 5000/yr (slightly below our usual rate), we would encounter a clinical candidate only every 20 yr of testing. However, let us propose that each of the three clinical candidates have 100 analogs each, of which 20 are moderately active (60 additional agents to detect). With totally random selection and testing of the inventory and the same 5000/yr testing rate, we would encounter 1 moderately active/yr. Each hit would lead us to test all the related analogs (which would contain the clinical candidate). Thus, it would take us only 3-4 yr to discover all three clinical candidates from this inventory." (Corbett et al. 2004, p.116-117)

probable one. Instead, it might be more desirable to have a screen that gives one candidate a year, albeit with a lower probability of being a true positive. As Lasagna mentioned, one can safely ignore the False negatives only provided that the screen anyway "produces a felicitous number of 'true positives' " (Lasagna 1958, p.939), which he however doubted would be the case. Nevertheless, as we will in the next section, sometimes increasing the throughput of the screen can be easier than increasing its sensitivity.

However, in some contexts the opposite argument might apply. For instance, in a discussion of animal models of psychiatric diseases, Paul Willner notes

> "In assessing the validity of a model as a simulation of the disorder, false positives and false negatives carry equal weight. However, in a screening test, these two types of error are not of equal importance. If a screening test accepts an ineffective compound (false positive), the error will eventually come to light in further testing and no permanent damage will have been done. However, if the test wrongly rejects an effective compound (false negative), a potentially beneficial drug will be lost irretrievably." (Willner 1991, p.8)

Willner's claim is too strong to be generalized to all situations, for in many circumstances trying an ineffective treatment is very damageable for the trial participants. Admittedly, this might be less common in psychiatrics, which is the field Willner is concerned with. Ultimately, the choice between the two strategies depends on the trade-off between the societal impact of discovering new drugs and that of testing False positives on human beings. Willner's argument applies especially in contexts where drugs are very scarce: a screen cannot afford to declare negative the only drug that might work. To some extent, the relative cost of false positives therefore depends on the expected frequency of effective drugs. However, the issue is not merely technical, for there might be moral reasons to protect patients and clinical trial participants even against long-term public health interests. Such deliberations are however beyond the scope of the present work.

2.3.2 The CCNSC – from attrition to annotation

At the NCI's creation by the 1937 National Cancer Institute Act, its drug discovery program was simply the continuation of federally funded research at Harvard University on sarcoma. The passage to a large-scale, systematic drug screening program was rather gradual, but the creation of the Cancer Chemotherapy National Service Center (CCNSC) in 1955 was a

key moment in the systematization of the drug discovery process.

The screen initially consisted of a primary screen followed by more diffuse secondary evaluations. The first stage was by far the most systematic, although the models changed over the years[9]; instead, the models used for secondary evaluation varied greatly and were rather chosen on a case-by-case basis.

The initial models were chosen on the basis of their coherence with clinical data, i.e. on their capacity to 're-discover' known chemotherapy agents – the Gellhorn-Hirschberg report mentioned earlier (Gellhorn and Hirschberg 1955). The models were induced murine cancers established as cell lines and transplanted into mice. For instance, the famous L1210 leukemia (the only model that was systematically used throughout the years) was developed in 1948 by painting the skin of mice with methylcholanthrene (Waud 2004). As the experimental procedure was inefficient and known to produce cancers that greatly differed from each other, the induction of the carcinogenic transformation was itself inappropriate for the purposes of a screen. However, transplanting cancer cells coming from the same cell population of these induced tumours promised to greatly increase both reproducibility and throughput.

Survival (or median increase in life span) was the first readout of the tests, and although it remained in the results (see for instance Schepartz et al. 1967), changes in tumour weight (or net log_{10} cell kill) quickly became the central assay. As data accumulated, it was clear that different models showed different sensitivity to compounds, and the screen was quickly perceived as biased towards certain forms of cancer (especially leukemia). In 1976, a panel of colon, breast and lung tumor models were introduced to complement the previous set of leukemia, sarcoma and melanoma (Waud 2004). Importantly, this panel also included xenografts from *human* tumors: the immunocompromised *Nude* mouse[10], which accepted human tissues, had been discovered in the 1960's (I will return to xenografts in the next chapters). However, because immunocompromised animals require very expensive infrastructures, their widespread use in a massive screening appeared unfeasible. And here

[9] "The initial CCNSC screen consisted of three mouse tumors: L1210 leukemia, SA-180, and mammary adenocarcinoma 755. Over the years, the primary screen changed from the original three tumors to L1210 plus two arbitrarily selected tumors to L1210 plus Walker 256 carcinosarcoma to L1210 plus the P388 leukemia to L1210 plus P388 plus B16 melanoma or Lewis lung tumor." (Waud 2004, p.80)

[10] Aside from its famous absence of hair, the nude mouse is characterized by its lack of a functional thymus and the corresponding massive reduction in T cells. As a consequence, it is largely unable to mount an immune rejection of foreign tissues.

again we see the same logic we saw earlier with the example of the Pap Smear and biopsy (section 2.2.2): xenografts on immunocompromised animals could be a right tool for the screen, but only in combination with some earlier screening step.

Up to the end of the 1970's, the screen produced very few anticancer drugs, and was generally perceived as a failure (Pagé 2004). In 1975, Vincent DeVita (who had occupied important positions at the NCI since 1963 and would a few years later become director of the NCI), reported to the director of the NCI that "The whole philosophy of screening needs reevaluation" (quoted in Keating and Cambrosio 2012, p.267). This general conclusion prompted a major change of strategy, the most important aspect of which was the adoption of an *in vitro* panel. Compounds would first be screened *in vitro* on a panel of cancer cell lines, and secondary evaluations would be tailored to each compound, according to its prior testing results and its known properties (Alley et al. 2004, p.127). Figure 2.1 shows a representation of the discovery pipeline.

What is important to note, here, is that *in vitro* cell lines are not the best models in terms of predictivity or similarity of response with respect to human tumours. On the contrary, cell lines are from a certain point of view extremely bad models because they are silent on a very important aspect of pharmacology: pharmacokinetics (e.g. where the drug goes in the body, how it gets modified, how fast it is degraded, etc.). Scientists were fully aware of this fact, and nevertheless chose to go *in vitro*. Their rationale was that, although one of the main advantages of *in vivo* models over *in vitro* models was that they purported to also model metabolism, the main difference between humans and rodents were metabolic. In other words, this meant that animal models, which were considerably more costly, gave little more information than *in vitro* screening, and introduced an additional risk of artefact in the form of metabolic asymmetries (see Pagé 2004). Furthermore, due to results in the early 1980's on colorimetry and automated counts, *in vitro* screening would dramatically increase the throughput, not only in the quantity of compounds screened (the new screen was planned for 10,000 substances/year[11]), but also the amount of information obtained from the screen. And it is this dimension that I would like to emphasize here: the passage to *in vitro* primary screening was not primarily meant to give better predictions

[11] To following passage gives an idea of the scale of the screen: "Using triplicate cultures for each concentration, five concentrations and 10,000 samples a year, the total number of culture units for the 60 cell lines was calculated to be 9,000,000 cultures a year." (Shoemaker 2006, p.815)

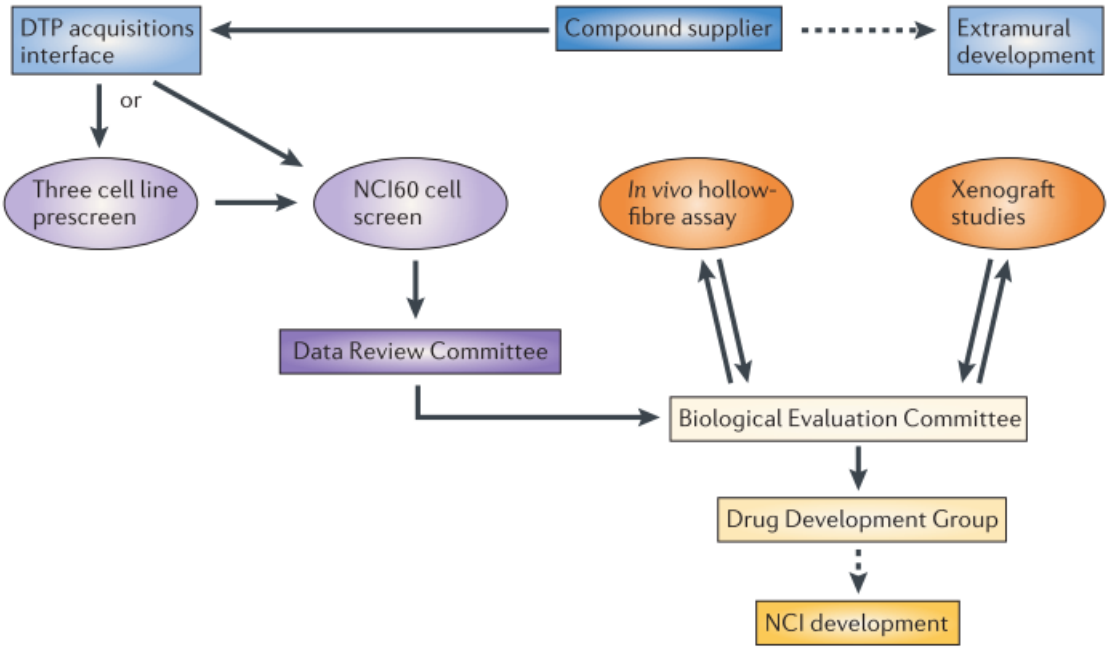

Figure 2.1: The CCNSC pipeline, adapted from Shoemaker (2006, p.817). The additional pre-screen composed of three cell lines was added in 1998 for synthetic compounds, further accelerating the discovery process..

than animal models, but instead represented an important shift in the very purpose of the screen. The models it relied on ceased to play the role of surrogate patients, and because observational instruments.

Michael R. Boyd, the main architect of the *in vitro* screen (and Associate Director of Developmental Therapeutic Program in the 1980's), describes its development in different places (Boyd and Paull 1995, Boyd 2004). In 1984, Boyd presented his project of a "disease-oriented in vitro primary anticancer drug screen" (Boyd 2004, p.42) to the Division of Cancer Treatment of the NCI, which was gradually set up and formally launched in 1990. The screen could be said to be 'disease-oriented' in two ways: it was not anymore centered on compounds, but on the relationships between compounds and cancer types; in a related way, it did not aim anymore at finding a one-size-fits-all drug against all cancers, but was prepared to single out drugs expected to be active on *some* cancer types. To this end, the screen relied on a panel of "60 different human tumor cell lines derived from seven cancer types including lung, colon, melanoma, renal, ovarian, brain, and leukemia" (Boyd 2004, p.47) – often referred to as the NCI60. Interestingly, the rationale for selection of the lines was purely a matter of internal validity (rather than external validity):

> "Selection of lines for inclusion in the panel required that they adequately

> met minimal quality assurance criteria (testing for mycoplasma, mitogen-activated protein [MAP], human isoenzyme, karyology, in vivo tumorigenicity), that they were adaptable to a single growth medium, and that they showed reproducible profiles for growth and drug sensitivity." (Boyd 2004, p.47)

For this reason, major cancer types such as breast and prostate cancer "were not represented in the initial panel because of unavailability of suitable lines" (Boyd 2004, p.46), and were added only in late 1992.

In a similar way, the intention was initially to include healthy cell in the screen to control for aspecific toxicity, this was in the end perceived as impossible:

> "The finding that normal cell types available at the time, that is, fibroblasts and certain epithelial cell populations such as renal epithelial cells, responded in vitro to anticancer drugs with extreme phenotypes (fibroblasts being pan-resistant and renal epithelial cells being pan-sensitive) in the assay selected for the screen led to the use of other tumour cell lines as controls." (Shoemaker 2006, p.813)

This is another important departure from a full analogy between the testing procedure and the expected clinical outcome: compounds were not assessed directly for their capacity to selectively kill cancer cells, but in a much more indirect way, through their profile of activity across different tumour types. It is important to note this discrepancy between what the test actually measures, and what this measurement is used for – an important feature to which I will return in the following chapters, and which is at the basis of the first assumption I wish to challenge: surrogacy (section 2.4.1).

At this point the screen did not have the purpose of simply rejecting or selecting compounds for further study, but instead aimed at providing different kinds of information on them:

> "Throughout, we have intentionally avoided characterizing compounds as 'active' or 'inactive' per se, or of suggesting any specific definitions thereof. This is to further emphasize the use of the NCI screen as a research tool, ideally to be employed in complement to diverse other screening, drug discovery and research strategies, rather than proposing any more absolute (and perhaps meaningless) activity criteria." (Boyd and Paull 1995, p.11)

In the most immediate way, the test outcome contained not a binary result, but a set of 60 qualitative and quantitative results: "The testing of a sample in the full 60-cell line screen generates a corresponding set of 60 dose-response curves." (Boyd 2004, p.48) Boyd and

Paull (1995) retrospectively writes of the screen as providing "a characteristic profile or 'fingerprint' of cellular response" (Boyd and Paull 1995, p.1-2). It is the whole pattern, instead of the result on any single cell line, which was the most useful. For instance, Shoemaker notes that "the patterns of relative drug sensitivity and resistance generated with standard anticancer drugs were rapidly found to reflect mechanisms of drug action." (Shoemaker 2006, p.813). And in fact, Boyd (2004) ends his paper with more recent developments going in the same direction:

> "To add new potential dimensions of utility, the NCI Developmental Therapeutics Program has continued to pursue the 'molecular characterization' of the cell lines of the NCI screen with respect to selected genes, gene products, and other possible 'molecular targets' contributing to maintenance or reversion of the malignant phenotype (see the DTP website at http://dtp.nci.nih.gov). It is hoped that such information will facilitate further use of the 60-cell screen and/or the accrued screening databases to discover novel molecular target-directed leads." (Boyd 2004, p.57)

Robert H. Shoemaker, who joined the Developmental Therapeutics Program (DTP) of the NCI in 1981 and is currently leading its Screening Technologies Branch, wrote that a 1997 review of the screen "resulted in a change in the mode of operation of the NCI60 from an NCI drug-discovery pipeline to a research tool in support of the cancer research community." (Shoemaker 2006, p.821) Instead, I would argue that the scientific discussion surrounding the screen shows that this change was a gradual one, taking place long before: at least from the beginning of the *in vitro* primary screen around 1990, and most likely much earlier. The focus on diseases, the fact that the screen lacked a binary or unidimensional result, and the format in which data were published and accessed, all point in this direction. Furthermore, data-mining algorithms specifically designed to analyze results of the screen, such as COMPARE (see Boyd and Paull 1995), were developed already at the onset of the *in vitro* screen (they would later be complemented by other algorithms such as self-organizing maps). The 1997 review was simply the recognition of this change.

A recent trend in high-throughput screening is to increase not only throughput, but also density: to increase the amount of assays performed (see for instance the workflow in Boisclair et al. 2004, section 6). In this context, biomedical models are not simply surrogates for human patients, but a panel of models is instead "used to profile the molecular properties of lead compounds" (Boisclair et al. 2004, p.27). Drug screening might have started as a

procedure of progressive exclusion of compounds, but it has become a procedure of annotation, in which assays are not simply sieves, but more complex observational instruments. In the next chapter, I will develop this view of biomedical models as instruments, which requires us to abandon some common assumptions regarding biomedical models.

While I have tried to emphasize the turn to a measurement of compounds, one must be careful not to consider the screen as 'compound-oriented', and keep in mind that the objects measured are not independent. First, as Keating and Cambrosio note, "it was often a question of choosing between compounds, rather than choosing a compound based on present criteria." (Keating and Cambrosio 2012, p.266) Moreover, as I mentioned earlier the screen is not random, because success with one compound recommends the testing of related compounds. There are families and sub-families of compounds (see Keating and Cambrosio 2012, p.248), and their exploration is itself an exploration of a space, in the sense that it teaches us something more than just what learn about the tested compounds.

2.4 Assumptions of the received views

The critique of animal models, my answers to it, and the example of the CCNSC screen raise important points regarding biomedical models, their use and their evaluation. In this section, I will try to unbundle a number of assumptions that have, in my opinion, hampered both our understanding and our evaluation of biomedical models. By and large, the philosophical literature on biomedical models has fallen prey to these assumptions, with the exception of some scholars who, questioning one, have tacitly endorsed the other.

2.4.1 Assumption 1: Biomedical models function as surrogates

Jessica Bolker has written that organisms can be used as models in two ways:

> "A model may be used either as an exemplar of a larger group, or as a surrogate for a specific target. [...] exemplary models most often serve basic research, while surrogate models are used when the target species we ultimately want to learn about is inaccessible or difficult to study, as in medical research." (Bolker 2009, p.485)

The same notion of surrogacy is present in the NRC's reports on biomedical models: "A biomedical model is a surrogate for a human being, or a human biologic system" (ILAR and NRC 1998, p.10) Biomedical models are indeed used in research projects for which, in most cases, human subjects would have been more suitable – were it not, of course, of ethical and economical limitations. This is not what I mean to deny when questioning the assumption that biomedical models function as surrogates. However, this abstract kind of surrogacy should not be meant to imply that the models are used *in the same way* human patients would have been used. Instead, what I will call the 'surrogate view' of biomedical models implies something stronger, namely that successful extrapolation implies an identity of outcome between the model and target systems. Bolker is again explicit on this point:

> "In carrying out research in surrogates with the intention of benefiting other species (particularly our own), we make a series of assumptions. One is that the surrogate will respond to manipulations in the same way as the target would, if it were examined directly." (Bolker 2009, p.490)

In such a view, a compound is deemed, for instance, toxic to humans on the basis of its toxicity to animals. The first thing to note, then, is the simplicity of the investigation: animals are apparently tested for the very same thing that one would test in humans, namely the toxicity of the exposure to a compound. Furthermore, the output of the model is a relational property between the organism (as a whole) and the compound. In this context, "the modelling relationships are reciprocal" (Committee on Models for Biomedical Research 1985, p.19) – humans could serve as models for the animal without significant change in the experimental design. As a friend and colleague of mine proposed (Fridolin Groß), what I take to be the 'surrogate view' of biomedical models can be summarized as 'taking whatever you can find that is the closest to human and then pretend that it is human'. I have several objections to this view, and believe that these characteristics are seldom encountered in biomedical models.

When a biomedical model is used as a surrogate, its modeling capacity is closely tied to its similarity to the target system. Even when arguments of common descent are used, they are but a proxy to similarity. This was clearly stated by the participants of the NRC workshops on biomedical models:

> "Models by homology are thus of heuristic value in the search for analogs,

> but they become functionally useful only when they are also good models by analogy for the phenomenon or structure being studied." (Committee on Models for Biomedical Research 1985, p.17)

Full identity between the model and target systems is obviously not required, and the similarity should instead be with respect to the features under investigation[12]. With respect to these relevant features, however, full recapitulation is generally said to be required. The model is for instance expected "to simulate human disease" (Piotrowska 2012, p.1), or to "reproduce the condition that existed in humans" (Greek and Shanks 2011, p.543). This is particularly obvious in the formalization used by Shanks et al. (2009): they construe the work of animal modelers as of going from the observation that "[i]n an animal test, X led to Y" to the prediction that "X will lead to Y in humans also" (Shanks et al. 2009, p.3), thereby implying that X and Y are the same thing in both systems. Likewise, in talking of "transferring causal generalizations" (Steel 2008, p.3) or asking whether 'the same effect' is present in both the model and the target system, most accounts of biomedical models seem to suppose an identity between the product of extrapolation (i.e. the knowledge sought about the target system), and what it is extrapolated from.

This issue was raised especially in the context of psychiatrics, where the problem is more acute. Willner (1991) argued that "There is no good reason to suppose that a given condition will manifest itself in identical ways in different species" (Willner 1991, p.14), and that as a consequence we should look for the corresponding – rather than identical – condition in the other species. He gives the following example:

> "For example, rearing on the hind legs is a prominent component of stimulant-induced stereotyped behavior in rats, but not in primates, whereas the reverse is true of scratching (Randrup and Munkvad, 1970). The physical topography of these behaviors is quite different; nevertheless, we are able to say that they are homologous across species." (Willner 1991, p.14)

As a consequence, models in this field are generally assessed in terms of *construct validity*, which is the basis on which Shelley (2010) proposed an account of extrapolation, as well as an answer to LaFollette and Shanks' critique. And this strategy has been adopted by

[12] In fact, Huber and Keuck (2013) emphasized that in many cases, complete recapitulation of a disease, for instance, might not even be desirable. The point has been raised by several scientists, especially in the field of Alzheimer's disease, who noted that "incomplete models of diseases may actually be advantageous" (Jucker 2010, p.1212). I will develop this point further in the next chapters.

some scientists: in a review on animal models for neurodegenerative diseases published in *Nature medicine*, Mathias Jucker writes that

> "If considered within the range of their validity, mouse models have been predictive of clinical outcome. Translational failure is less the result of the incomplete nature of the models than of inadequate preclinical studies and misinterpretation of the models." (Jucker 2010, p.1210)

Interpretation is the key term here, and I will discuss this point in more depth in Chapter 4. My argument against surrogacy, however, goes further than Shelley's replacement of identity with correspondence. As the example of the CCNSC screen shows, the screen went from looking at overall survival to looking at the weight of tumours, before changing to the 'profile of activity' of the compounds – something quite remote from curing patients. To understand how this can be used for drug discovery, one must abandon the 'surrogate view' of biomedical models.

2.4.2 Assumption 2: Modeling is a unidirectional process

Most accounts of modeling seem to assume that things are first worked out in the models, before being extrapolated to the target system. This is for instance a distinctive feature of modeling on Weisberg's (2007) account: analysis of the model and extrapolation come in two distinct stages of modeling[13]. This same unidirectionality is generally transposed to the process of translational research. Indeed, the very idea of translational medicine already implies a distinction between two systems, a directionality, and some pre-existing fact or knowledge, already there to be translated. In his editorial inaugurating the creation of the Journal of Translational Medicine, the editor in chief emphasized that translation is "a two-way road" (Marincola 2003). Upon closer inspection, however, one notices that behind the words persists a strong unidirectional conception, and in fact the other direction – 'Bedside to Bench' – seems limited to the provision of (indeed precious) *ex vivo* materials. Talking of a two-way road therefore acknowledges a bidirectional transfer, but each direction being quite restricted and different from the other: predictions on one way, raw materials on the

[13] "In the first stage, a theorist constructs a model. In the second, she analyzes, refines, and further articulates the properties and dynamics of the model. Finally, in the third stage, she assesses the relationship between the model and the world if such an assessment is appropriate." (Weisberg 2007, p.209)

other. This is mistaken, both descriptively and normatively[14]. Not only the directionality is problematic, but the very idea of linearity: as DeVita said regarding the NCI's screen, "[s]ome compounds may actually be at several stages at the same time." (quoted in Keating and Cambrosio 2012, p.265-266)

One point that I particularly wish to emphasize is that the clinical can also provide clinical meaning to a phenomenon in the model, and an example from Nicole Nelson (2012) illustrates this very simply. Nelson made a similar criticism of unidirectionality in her study of the scientific usage of the elevated plus maze. The maze is the basis for rodent behavioral models of anxiety. It consists of an elevated, plus-shaped maze with two open arms (without walls) and two closed arms (surrounded by high walls). The proportion of time the animal spends in the closed arms is interpreted as a proxy to, or measurement of, its anxiety. Nelson writes:

> "Describing animal models as 'exemplars' or modeling as a process of 'extrapolation' creates the impression that findings are first worked out in model systems and then translated to other cases, thereby obscuring the ongoing interactions between the model and the modeled that take place during all stages of animal modeling work." (Nelson 2012, p.6)

Perhaps the clearest example of such interactions is what she calls the "pharmacological argument": namely the prediction, from the clinically established fact that a drug (*Librium*) relieves anxiety in humans, that it should also relieve the murine phenotype interpreted as anxiety (Nelson 2012, p.11). The point is well known to scientists, especially in the field of psychiatrics: "Drugs are vital for characterizing models that, in turn, are essential for evaluating actions of drugs" (Millan 2008, p.5). As Nelson points out, once this clinical prediction is corroborated in the model, it serves to establish the credibility and clinical relevance of the model. Huber and Keuck describe this as a bidirectional approach including "processes of generating" and "strategies of validating" experimental organisms (Huber and Keuck 2013, p.386-387). It must be emphasized that this is not a sequential, but an iterative process. Furthermore, casting it in terms of 'validation' underestimates the value of bedside-to-bench extrapolations. Instead, I believe they do much more: they also

[14] The point was summarized in the simplest way by Frederic Rosa in a conference on zebrafish models in translational medicine: unlike us, he noted, the fish will never go to the physician to complain that it is ill; clinical cases can reveal things in the model which would otherwise go unnoticed (Frederic Rosa, closing remark, at the conference *Zebrafish Models in Translational Medicine*; Courcelle-sur-Yvette, France, 4th of October 2013).

establishes benchmarks according to which variations in phenotypes are evaluated, thereby defining, in Nelson's example, the very phenomenon of murine anxiety.

2.4.3 Assumption 3: Modeling is a dyadic relationship between a model and a target

Much of the discussion on models – not only biomedical models, but scientific models in general – has focused on the relationship between a 'model-target dyad'. This dyadic view of models has been criticized in that it ignores the modeler and his/her aims (see Giere 2004, Knuuttila 2011). My criticism is however of a very different nature, for I want to argue that models do not stand on their own in the economy of the lab. Modeling, I will argue, is best understood not as a relationship between a model and a target system, but as a process distributed in many inter-related systems.

Once more, this is illustrated by the CCNSC screen: in the early phase of the CCNSC, the main reason why both the P388 and L1210 leukemia were used is that they had different sensitivities to difference active agents[15]. Supposing that there had been a model which was more predictive than any of the two taken in isolation, it might still have been less useful. The reason is that two weak but complementary models might be better than stronger (but still imperfect) models. Different models can correct each other's weaknesses, although articulating them in such a way is not always straightforward. The point, here, is that each model ought not be evaluated on its own.

These cases are very simple, for although the value of the model cannot be assessed in isolation, it can still function as a model in the same way when taken in isolation (and hence counts as a model according to the criterion proposed in section 1.3.1). My claim is however stronger: I will argue that often, the modeling relationship exists only (or is transformed) in its conjunction with other model systems. For the moment, it is sufficient to observe that the *in vitro* screen of the CCNSC already goes in this direction: since the primary information provided by the NCI60 panel is a 'profile of activity', the value of the readout of any of the cell line depends on the activity in the others. This is all the more relevant if one considers all the other assays of the CCNSC screen which I have not

[15] For instance: "The vinca alkaloids are active against P388 leukemia but ineffective against L1210 leukemia". In turn, L1210 is more sensitive to hydroxyurea (Waud 2004, p.81).

discussed here (hollow fiber assays, colony-formation assays, etc.).

2.4.4 A new account of biomedical models

By and large, traditional accounts of biomedical models have relied on a dyadic and unidirectional view of modeling, in which model systems function in isolation, as interchangeable surrogates for humans. This view has major shortcomings, perhaps the most important being that it leaves unanalyzed the relationship between models, and thereby fails to account for the interdependence of different model-target relationships. In the following chapters, I will propose an alternative understanding of the way biomedical models function and which emphasizes they synergy.

Most criticisms of animal experimentation have missed their target precisely because they have criticized a narrow view which is inadequate to the way biomedical models are actually used. In a certain way, the transformation of the CCNSC screen can be seen as the victory of Shanks' and Greek's arguments against a strictly predictive usage of animal models. However, the fact that this shift occured in the early 1980's (and was already intensely discussed in the early 1970's), in other words long before Shanks and Greek (2009) or LaFollette and Shanks (1996), shows that the criticisms were off-target.

Biomedical models are experimental systems (Chapter 1), and as such must be understood through the 'niche' they occupy in the ecology of scientific research (Rheinberger 1997, p.227). As a consequence, a proper epistemological understanding of biomedical models and modeling necessarily involves a philosophy of experimentation. For this reason, the next chapters will pay a much closer attention to experimental practice. The general method will be one of analysis and synthesis. In Chapter 3, I will try to distinguish different manners in which biomedical models are used in the lab – different proximal functions of biomedical models. Challenging the first assumption (surrogacy), I will focus on models that function as measuring or detection devices. In Chapter 4, I will explore their relationship with the theoretical realm. Finally, in Chapter 5 I will then consider how the models work together, dealing more directly with the last two assumptions.

Chapter 3

The instrumental role of biomedical models

3.1 Introduction

In Chapter 1, I have shown that applying the concept of model to biomedical models is problematic, especially with respect to the distinction between modeling and what we might call 'direct experimentation', leading to a conflation of models (as can be applied to biomedical models) with experimental systems. Most of experimental biology, I suggested, is modeling[1]. It follows that modeling should not be understood in opposition to experimentation, but instead that a philosophy of biomedical models should at the same time be a philosophy of experimental biology, and here I will attempt to bring the two together.

In Chapter 2, I have argued that a simplistic understanding of biomedical models has prevented their proper assessment. I have highlighted three problematic assumptions of mainstream views about biomedical models (section 2.4), and in this chapter I will further my criticism of the first: the assumption that biomedical models function as surrogates for human patients (section 2.4.1). The surrogacy view assumes that the model is used in the way the target system (a human subject) would have been used, were it not for moral and practical limitations. When a biomedical model is used in this way, it is a good surrogate to the extent that it is a replica of the target system: although the model is

[1] In the philosophy of economics, Mäki (2005) has also argued that models are experiments, and that experimentation is modeling.

more manageable, extrapolation depends on the model being similar to the target system in all relevant aspects. There is no doubt that biomedical models are sometimes used in this way; my claim here will instead be that they are not always used in this way, and in fact I would argue that it is most often not the case. To show this, I will distinguish from the surrogate role of biomedical models a radically different role: the instrumental role of biomedical models. After a brief discussions on the kinds of distinctions that can be made about biomedical models (section 3.1.1), I will first illustrate the instrumental role with some examples (section 3.2). I will then distinguish different kinds of scientific instruments and attempt to provide a definition of the two roles in section 3.3.

The purpose of this discussion is two-fold. First, because the instrumental role is so different from the surrogate role, it suggests a different, and I would argue more fruitful way of looking at biomedical models. Second, characterizing the distinction between the surrogate and instrumental roles can draw attention to very important dimensions according to which biomedical models can differ in the way they contribute to biomedical research. For if in the long run biomedical models are expected to inform us about human pathologies, *they get there through a wide variety of ways*, and their evaluation must take these differences into account.

3.1.1 Kinds of kinds of biomedical models

There are obviously many ways to draw distinctions among biomedical models, and such distinctions can either focus on the model systems themselves, or on the way they are used. The distinction between *in vivo* and *in vitro* models, which will be discussed at length in Chapter 5, is an example of the first kind – what we might call structural distinctions. Many more such distinctions can be made, although not all of them are robust or useful[2], and I would argue that most of them are taken as a proxy for what is supposed to map between the model and the target.

A more interesting example is the distinction between positive and negative disease models, the latter being understood as models that lack the phenomena of interest. Two interesting examples include mouse strains resistant to the parasitic disease Leishmaniasis

[2] The distinction between spontaneous and induced models (models which develop a disease either 'naturally' or because of specific interventions) is an example of a particularly accidental distinction.

(Farah et al. 2002), and animals displaying a particular resistance to cancer, such as naked mole rats (see Tian et al. 2013). Most often, experimentation on negative models aims at triggering the phenomenon, for such interventions shed light on their mechanisms of resistance. These are of potential biomedical relevance not only because they might suggest therapeutic strategies, but also because they shed light on pathogenesis and the conditions it requires[3]. But they do so only at the moment when they have finally tempered with so as to develop the disease, and the difference with positive models is merely as to which of the two – diseased or treated – comes first.

Here, I will instead concentrate on the other kind of distinctions one can make among biomedical models, namely those focusing on the different ways models are used. I believe they are much more relevant to epistemological analysis than structural distinctions. Once more, Shanks and Greek provide an interesting starting point for this discussion:

> "Animals are used in science in at least nine distinct ways: (1) as predictive models for human diseases; (2) as predictive models to evaluate human exposure safety in the context of pharmacology and toxicology (e.g., in drug testing); (3) as sources of 'spare parts' (e.g., aortic valve replacements for humans); (4) as bioreactors (e.g., as factories for the production of insulin, or monoclonal antibodies, or the fruits of genetic engineering); (5) as sources of tissue in order to study basic physiological principles; (6) for dissection and study in education and medical training; (7) as heuristic devices to prompt new biological/biomedical hypotheses; (8) for the benefit of other nonhuman animals; and (9) for the pursuit of scientific knowledge in and of itself." (Shanks and Greek 2009, p.30)

Shanks and Greek rightly point out differences in the way biomedical models (animal models in their case) are used, but their list is heterogeneous to say the least. Some items on the list are concrete products of the systems (e.g. producing materials such as antibodies or insulin), whereas other items are defined by the kind of inferences in which they are used (e.g. prediction or heuristics) or the particular meaning which such an output takes in our society (e.g. evaluating safety and therapeutic effect). As a consequence, one can expect a lot of overlap between these elements. A first clarifying step is to distinguish between what we might call the *proximate and ultimate functions* of biomedical models. Proximate

[3] Importantly, what makes the naked mole rat an interesting model *for* cancer is that they are in other respects relatively good models *of* humans, in the sense that many of their features map onto features of human physiology. By contrasts, the fact that plants do not suffer from cancer is not sufficient to make them good (negative) models for cancer.

functions are the immediate usage of the model in the lab, for instance the kind of output it produces. Ultimate functions, or goals, instead represent the contribution the model makes to the broader goals of the scientific enterprise in which it participates. Some models are used for regulatory purposes – for instance to know whether (and up to what concentration) certain chemical compounds should be allowed in marketed products. Other models are for drug discovery, or with a variety of other goals in mind: explanatory, exploratory, etc. The same model (in the sense of the same experimental system) can be used for different such purposes, in which case it changes its ultimate functions. However, it does not necessarily change its proximate function – the way it is actually used in the lab. Conversely, even with the same ultimate goal in mind, there are different ways in which a given model can be used to advance this goal – different proximate functions.

As discussed in the previous chapter, the ultimate goal of a model makes an important difference to the way it is evaluated, notably to the way sensitivity and specificity are valued (see especially sections 2.2.4 and 2.3.1). In this chapter, I will deal mostly with the proximate functions of biomedical models. Among these, I will focus on what I believe is the most important distinction, which was already illustrated in the previous chapter in the shift of the CCNSC screen from using models as surrogates to using them as observational instruments (section 2.3.2). As previously emphasized (section 2.4.1), surrogacy can take different meanings when it applies at different levels: biomedical models are by definition used in research projects for which, in most cases, human subjects would have been more suitable, were it not, of course, of ethical and economical limitations. This fact, however, does not imply that these organisms are themselves surrogates for humans: that they are used instead of humans in experimental designs which are otherwise roughly the same. An important point of this chapter is to show that biomedical models can be informative about human biology without being surrogates in this sense.

3.2 Living instruments

In this section, I will suggest that biomedical models are often used as instruments, and propose a characterization of this role. The idea of seeing animal models, especially model

Important parts of this section have been previously published in Germain (forthcoming a).

organisms, as instruments is not new, and appears in a number of studies in history and philosophy of science (Kohler 1991, Rheinberger 1997, Rader 2004, Landecker 2007, Gayon 2006, Gachelin 2006). However, as Marcel Weber remarked (Weber 2005, p.169-173), it would seem from these accounts that 'instrument' is used in a rather superficial, metaphorical way, or that the analogy refers mostly to the social transactions involving model organisms: indeed their standardization is, from a social and historical point of view, very reminiscent of that of measuring devices. If the notion of instrument is to be more than a metaphor, or at least a useful metaphor, the analogy has to extend further into the epistemological realm. Weber, in his criticism of this view, lays out the criteria for a proper (more than metaphorical) analogy would have to fulfill:

> "In a typical instrument, a causal input (e.g., the Earth's magnetic field in the case of the compass) leads to an observable signal (the motion of the compass needle). What the user of an instrument is interested in is the process or object that is responsible for the causal input (the orientation of the Earth's magnetic field). [...] It seems to me that, rather than using an experimental organism to detect some causal input, the experimental biologist is interested in the organism itself. To be precise, the experimental biologist wants to understand the biological and physicochemical processes that occur in the organism. The laboratory organism is not the measurement or observation device; it is the object that is being measured or observed." (Weber 2005, p.170-171)

It is important to note here that Weber understands 'instrument' in a narrower sense than we commonly use it: a scalpel is clearly an instrument (made to fulfill a function), but does not seem to fit Weber's description (it does not transmit a signal). The instrument Weber has in mind, here, is of a more precise sort: the observational instrument. I will distinguish different kinds of instruments in section 3.3.1, but for the moment it is also in Weber's sense that I will use the word.

One can distinguish several kinds of observational instruments, and I will suggest that in order to talk of a measuring or detection device, there needs to be a kind of decoupling between the output of the device (the signal) and what this allows us to infer in the input or target system (temperature is not the same as the height of a column of mercury). Among this subset, I will consider those that give a binary output (yes/no) detection devices, while a measuring device should at least provide ordinal, if not quantitative readouts. In what follows, I will refer to both measuring devices as the 'instrumental role' of models.

It is uncontroversial that laboratory organisms are sometimes used in this way. Weber gives the example of animals brought into mines to detect toxic gases[4]. In a similar way, when Boyle put a mouse in his pump and saw it die, he was not trying to study the mouse, but was rather using it to learn something about the environment inside the pump (he was giving a shot at the Herculean task of detecting emptiness). The relevant question, therefore, is not whether there exist examples of models as instruments; it is rather about the importance of this instrumental role in understanding biomedical research.

In order to show the relevance of this role, I will start with some of the clearest examples, and gradually work my way to less obvious ones. I will begin with an example of an animal case which very clearly instantiate an instrumental role of biomedical models (section 3.2.1), followed by an *in vitro* model which is also best considered as a measurement rather than as a model (section 3.2.2). I will then turn to two additional examples (sections 3.2.3 and 3.2.4) showing that the instrumental role is much more prevalent than it might seem and applies to much more conventional types of disease modeling. On the basis of these examples, I will later propose a more precise characterization of the instrumental role (section 3.3), to be distinguished from the surrogate role.

3.2.1 Example 1: The Ascheim-Zondek test for pregnancy

In the 1920's, Bernhard Zondek and Selmar Ascheim developed a pregnancy test – the Ascheim-Zondek test, or A-Z test (see Zondek 1928) – very similar in functioning to the tests that can be bought at the drugstore. In this test, juvenile mice are injected with the urine of a female patient, and after two days are dissected. If the injection caused a maturation of the mouse's ovarian follicles, then the woman is pregnant. The mouse allows one to detect, in the input, something that was otherwise unobservable.

The rationale for this method was grounded on the conservation of the hormone governing the maturation of the follicles in both species (Zondek 1928, p.1088), and therefore on

[4] A very similar strategy was also proposed in a scientific context: in his contribution to the second NRC meeting on 'Animal Models for Biomedical Research', Walter E. Brewer discusses the use of animal species as indicators of contamination of ecosystems. He draws a hierarchy of organisms according to sensitivity ("In general, mayflies are the most sensitive organisms in a stream" ILAR and NRC 1969, p.19), and proposes a measurement of toxic materials in an ecosystem on the basis of population histograms. Like in the case of the *in vitro* screen of the CCNSC, what is interesting here is that each organism (or in fact each species) is by itself uninformative, and it is taken together that they provide information about the ecosystem.

a partial similarity between humans and mice. Nevertheless, the mature design of the test presented in the 1928 paper is clearly the result of an induction largely independent of questions of similarity. Indeed, the authors measured the accuracy of their test on a number of women in various conditions before concluding that the swelling of the reproductive organs is not a trustworthy signal: it can occur in the mouse, in different situations, even if the woman is not pregnant. Rather, they concluded that the relevant signal was the presence of small blood spots (the "Blutpunkte" of Reaktion II) on the follicles (Zondek 1928, p.1089), a phenomenon whose presence or absence in the pregnant woman is not even discussed in the paper. The reason is simple: although similarity between the two species might have prompted the test in the first place, its instrumentalization only requires that the urine of pregnant women gives a a reproducible signature in the mouse, independently of the nature of this signature.

The mouse, therefore, is not used as a surrogate or replica of the patient. Rather, it seems justified to talk of a measuring (or at least detection) device, as the animal clearly has the function of detecting a signal in order to learn something about the woman. While the A-Z test could easily pass for an historical oddity, I wish to suggest that the instrumental role is actually very relevant in contemporary biomedical research. In fact, the most interesting aspect of the A-Z test is its striking similarity to a very common model in cancer research: xenograft models, to which I will turn in section 3.2.3.

3.2.2 Example 2: Induced pluripotent stem cell models of cancer

A new kind of cellular models has recently made its appearance with the emergence of the technology of cellular reprogramming. As this will be discussed at length in Chapter 5, here I will only provide the necessary basics to understand the example. This technology allows one to transform ('reprogram'), through the transient expression of key transcription factors, differentiated cells (typically skin fibroblasts) back to a state of pluripotency, termed induced pluripotent stem cells (iPSC), which is very similar to that of embryonic stem cells. It is then possible to differentiate these cells into a tissue completely different from that of origin, by channelling their development with specific cocktails of transcription factors. This is particularly promising for the study of neurological diseases where we cannot, for

the sake of research, simply harvest neurons from patients. As I will discuss in Chapter 5, the process of reprogramming involves additional biases which need to be controlled, but in many circumstances, it is by far the best modeling method available.

What can be particularly puzzling, however, is the use of iPSC when other cells are readily available that are much more similar to the target system. An interesting example is the cellular reprogramming of cancer cell lines (see for instance Carette et al. 2010, Kim et al. 2013; for a review, see Suvà et al. 2013). Cancer is studied with a variety of models, including *in vivo* models and cell lines derived from patients. Although access to biopsies is limited (and indeed a very political matter), cultured cell lines are available for most cancers. Despite their limitations, cell lines are certainly more similar (whatever the way similarity is assessed) to the original tumour cells than the same cell lines reprogrammed to a pluripotent state. Why, then, use iPSC to study cancer?

There are a number of reasons[5], but the most important use was to detect something very specific about cellular phenotypes. By definition[6], cellular reprogramming erases the epigenetic memory of the cell, therefore allowing us to distinguish the genetic from the epigenetic. If a phenotype of the cancer cell is still visible in the iPSC (or iPSC-derived cells), then chances are that it is mostly genetic. If, however, it disappears in the iPSC, then it has epigenetic determinants[7].

It should be clear, here, that the iPSC model is not a surrogate for the original tumour, for the un-reprogrammed cancer line would be a better surrogate. Like the example of the caricature presented in Chapter 1 (section 1.1.3), iPSC models of cancer are *in all respects* less similar to their target system than the cancer cells lines from which they are

[5] Among the cancer cells' many genetic changes, most are relatively neutral. But they are neutral in the micro-environment in which they were acquired: when the cell is in a completely different epigenetic state, these changes might turn out to be lethal. This can also be true of driving oncogenic mutations: c-myc hyperactivation, for instance, is oncogenic (and therefore very fit from the point of view of the cell, because it induces proliferation) in some cell types, but lethal in others. So part of what researchers want to see is whether these cells could be reprogrammed at all, and whether they could go again through differentiation paths.

[6] The adequacy of this definition is still a controversial issue, and there have been debated reports according to which reprogrammed cells retain an epigenetic memory of their tissue of origin (Kim et al. 2010). Furthermore, as a recent persepective has emphasized (Ladewig et al. 2013), speaking of 'erasing' the epigenetic makeup of the cell gives the wrong impression that pluripotent cells lack any epigenetic makeup, when in fact their simply have a different profile. Nevertheless, reprogramming involves such drastic changes in the epigenetic profile of the cell that it can still be used as discriminating device.

[7] One can even further divide the effects by re-differentiating the cells into the original lineage: if the feature reappears, then it is not only epigenetic, but associated to the developmental state, while if it does not reappear, it is has epigenetic determinants which are specific to carcinogenesis.

derived. Yet like the caricature, they are used because they are particularly suited for a very specific kind of enquiry. More specifically, the iPSC model is an instrument to provide a very specific kind of information – to detect differences that were otherwise hidden. But like the example of the 'activity profile' of the CCNSC screen (section 2.3.2), the value of the readout provided by the iPSC model rests on its comparison with the non-reprogrammed line.

A natural reaction to this example would be to argue that this is more a case of experimentation on the original cells than a model. As noted in Chapter 1, however, drawing a distinction between modeling and non-modeling (direct experimentation) is particularly problematic in the context of biomedical models. Indeed, an experimental system is always different from the scientific object it is meant to teach us about[8], and the more so when studying entities that are not stable over time and that change according to their environment. Instead of taking for granted a problematic division between modeling and direct experimentation, I propose to seek other – and perhaps epistemologically more fruitful – parameters, such as the distinction that I will draw between instrumental and surrogate roles.

3.2.3 Example 3: Xenograft models of cancer

Following the nomenclature of Snell (1964), a xenograft (or xenotransplantation) is a case of tissue transplantation where the donor and recipient are of two different species. From the end of the 19th to the middle of the 20th centuries, transplantation was a topic of scientific fascination, and scientists attempted a vast diversity of transplant experiments. This was especially common in the field of cancer, in an attempt to domesticate tumours to the laboratory. Human tumours, if they were to be studied experimentally, needed to be studied outside their host. Even in the case of animal tumours, scientists were confronted with the simultaneous shortage of spontaneous tumours and inability to sustain a tumour beyond the death of its host. Hence transplantation became (and still is today, although for different reasons) among the most widespread ways of studying cancer in a lab.

[8] "The experimental conditions 'contain' the scientific objects in the double sense of this expression: they embed them, and through that very embracement, they restrict and constrain them." (Rheinberger 1997, 29)

There is some disagreement as to the first author to be credited with successful tumour transplantation. Claims go back at least to 1889 (Mayet 1902, Ewing 1919), and perhaps a couple of decades earlier, but were all strongly criticized – see for instance Hekzog (1902), who instead credited the feat to Leo Loeb. More recently, a historical review of chemotherapy attributes the "first transplantable tumor systems in rodents" to G.H.A. Clowes in the early 1910's (DeVita and Chu 2008, p.8643). The contention seems to hinge on what is stable enough to constitute a 'system'. Indeed, an important reason for the disagreement is that for a long time, the criteria on which to evaluate a successful graft were unclear (see Loeb 1945, chapter 12). In general, grafts lasted only for some time before resorbing under the pressure of the host's immune system[9]. Therefore, a line had to be drawn somewhere to distinguish cells that have successfully engrafted, albeit only temporarily, and cells that are just 'still there' from the injection. Some authors were already discussing histological criteria, for instance vascularization, but for a long time there was no established way to make the distinction. A related problem is that the injection caused an injury to the recipient that had important risks of infections, which (either because of the inflammation or of the death it brought) could easily pass for cancer.

Carl O. Jensen reported in 1902 the reproducible growth of a murine cancer through several generation of serial transplantation, but noted that this was highly dependent on the strain of the mouse (Woglom 1913). The systematic, large-scale work of Leo Loeb (especially from 1901 to 1910) was certainly of central importance in the establishment of transplantation systems (Witkowski 1983), but xeno-transplantations were of limited success for a long time (see for instance Funk 1915). The main improvement in this respect came from the discovery that some locations in the host (the brain, the anterior chamber of the eye, etc) accepted grafts more readily than others. As grafts started to become more efficient, and transplantation systems were tamed, the possibility appeared of using them as tools for a variety of purposes. The long established observation that only embryonic and cancer tissues were transplantable across species (normal adult tissues 'did not take') lead to such instrumental uses. Some scientists, for instance Harry S.N. Greene[10], proposed that "transplantability constitute[s] a biological test for cancer", and

[9] At that time, the explanation was that the foreign cells lacked specific 'foods' – it was not until Peter Medawar's work in the 1940's that the immunological basis of rejection was firmly established.

[10] Greene (1904-1969) was chairman of the department of pathology at Yale University, and was recog-

suggested that the "study of the transplants allows a more precise classification than is warranted from the morphologic features of the biopsy specimen" (Greene 1948, p.1364). Greene was explicitly proposing a diagnostic tool to replace what he considered to be a 'coarse' and uninformed judgement by pathologists.

Importantly, transplantation was not simply believed to be a useful signal: if it was a good signal, it was because it was signalling *something*, and therefore giving access to some invisible differences between cancer cells:

> "The fact that a biological quality as fundamental as the ability to grow in an alien species differentiates morphologically identical tumors suggests that the tumors must also differ in metabolic or biochemical constitution. It would seem important, therefore, to distinguish tumors with respect to this property and to study the different groups formed rather than to consider morphological similarity a proof of constitutional identity." (Greene 1952, p.41)

The very idea of using transplantation as a test implied that transplantation made visible a difference that was already in the tissues. More importantly, transplantability was not understood as binary: degrees of transplantability could be obtained either by resorting to statistics (the proportion of cases where the transplant was successful) or by assessing the pace, duration, and quality of the growth. Hence more than a tool to detect malignancy, transplantation was a tool to measure it. Indeed, as I will detail in Chapter 4 (especially section 4.3.1), these transplantation systems led to the invention of abstract quantities thought to be possessed by cancer cells. While it is arguable whether such quantities were numerical in a strong sense, they were at least ordinal: by 1952, Greene had ranked over one hundred tumours on this basis.

The use of transplantation systems to assess the malignancy of tumours of cancer cells was still very important in the discussions of the Cancer Chemotherapy National Service Center (CCNSC) at the end of the 1950s (see for instance the workshop discussion transcripts in the Annals of the New York Academy of Sciences, 1958), and one could argue that it never disappeared. Indeed contemporary xenograft experiments, especially in the field of cancer stem cells, are surprisingly similar to Greene's test. As I will discuss in detail in Chapter 4 (section 4.3), xenotransplantation is commonly used nowadays for the

nized internationally for his work on tissue transplantation, especially his techniques for tumor transplantation.

detection and identification of Cancer Stem Cells (CSC) – the alleged minority of cancer cells responsible for sustained tumour growth. The basic strategy is to divide the population of tumour cells into subpopulations according to some markers, and assess whether and to what extent these subpopulations contain CSCs by serially transplanting them into immunodeficient mice. Scientists first transplant into a first mouse, harvest the cells of the newly grown tumour, and transplant them into a second mouse. Roughly put, if the second mouse develops tumours, then the initial population contained CSCs (obviously, practice is fraught with challenges and complexities, some of which I will explore in the next chapter). In this context, the mouse is therefore an instrument for the detection of CSC.

The use of immuno-compromised mice is meant to reduce immune reaction to the foreign (trans-specific) tissue, otherwise one could not make the difference between an immune rejection of the transplantation and an intrinsic incapacity of the cells to form tumours. The idea here is not that the immuno-compromised mouse is more similar to humans – it is not, for very few cancer patients lack a functional immune system. As some colleagues noted, "the impairment of the immune system in NOD/SCID mice does not make these creatures more human but rather less murine." (Maugeri and Blasimme 2011, 612). The focus on removing the murine specificity rather than mimicking human specificity is best understood as controlling errors (Weber 2005), or more specifically controlling for artifacts rather than striving for a greater similarity.

Other xenograft experiments have similar instrumental roles. For instance, Topczewska et al. (2006) transplanted human melanoma cells into zebrafish embryos at the blastula stage, and noticed ectopic structures on the head of the developed larvae which were strongly reminiscent of "the duplicated axis formation induced by the dorsal organizer in classical dorsal lip transplantation studies" (Topczewska et al. 2006, p.925), particularly by the morphogen Nodal:

> "By acting as an organizing signal before gastrulation, Nodal initiates embryonic axis formation, and previous studies have shown that ectopic expression of Nodal induces mesendodermal fates in ectopic positions. Given the similarities between the axes initiated by the aggressive melanoma cells and those organized by ectopic Nodal expression, we considered that aggressive melanoma cells might secrete Nodal, thereby influencing zebrafish development." (Topczewska et al. 2006, p.926)

Indeed, their paper published on *Nature medicine* reports the role of Nodal in melanoma invasiveness. The authors write that they have "used embryonic models as a tool to discover molecular mechanisms by which cancer cells modulate their microenvironment" (Topczewska et al. 2006, p.925). Commentators have suggested that using the zebrafish embryo "as an *in vivo* biosensor for factors that are produced by the tumour" represents "a unique assay" (White et al. 2013, p.628). Admittedly, to call this a biosensor is far-fetched, and the authors note that this has not yet been performed for any other factor. Nevertheless, there are several elements which make the zebrafish particularly amenable to such an approach. Because the zebrafish embryo is easily produced (large clutch sizes) and easily studied (the embryo is transparent and accessible), a considerable amount of knowledge and experimental results have been generated regarding its early development. This means that there is a large repertoire of phenotypes to be recognized and to be compared to. In other words, standardized and systematic phenotyping of large-scale repositories, such as the genome-wide knockout mice (an ongoing project of the Sanger Institute for the generation of mice mutants for each gene of the genome, see White et al. 2013), represents more than a study of the functions of each of these genes: it also provides a map on which it will be possible to locate phenomena observed in future experiments.

It is important to note that although both animal clearly have an instrumental role in the experimental paradigm presented, this is not intrinsic to the systems and instead comes from the way they are used. Indeed, the very same xenograft models can also be used as surrogates for humans: if a researcher studies the way the transplanted tumours prompt the formation of blood vessels in the neighbouring mouse tissues, and infers that a similar process of angiogenesis occurs in humans, the xenograft model is expected to replica human phenomena. When this is the case, the animal does not simply has the aim of making visible unobservable differences, but has the additional aim of mimicking the human inner milieu. Just like the injected cells *stand for* the cells of the human tumour, so does the mouse stand for the original tumour environment.

3.2.4 Example 4: Genetically engineered models of cancer

The xenograft experiments mentioned here and the A-Z test seem to share a relevant characteristic allegedly absent from most cases of animal models: in both cases, the model includes a material input from the target system. It might be thought, therefore, that this is what characterizes the instrumental role from more traditional biomedical models. In fact, however, this material transfer is neither sufficient, nor necessary, for the instrumental role. It is not sufficient because, as mentioned in the previous section, the same material system can be used as both an instrument and a replica of a human being. Furthermore, as I argued in Chapter 1 the model is not limited to the animal, and very often some elements of the model system are indeed transferred from the target (i.e. clinical) system. In this section, I will rely on an example of another (and much more common) kind of animal model to show that this material transfer is not necessary either for the instrumental role, and thereby show that the instrumental role applies to a much broader set of animal models.

Santoriello et al. (2010) developed a genetically engineered zebrafish which, due to the over-expression of an oncogene (HRAS) selectively in melanocytes, developed melanoma regularly after some weeks. Like many oncogenes, genes of the RAS family were first discovered in a transforming virus before a homolog was found in human cells. HRAS mutations, as well as the specific mutation used (glycine-to-valine at position 12), are very frequent in human cancer, especially in skin cancers. It was therefore to be expected that the fish would develop a cancer akin to human melanoma. Strictly speaking, the model should not be considered a model of 'melanoma', but a tool in the investigation of HRAS-mutated tumours of melanocytic origin.

Like the xenograft example, this genetically engineered model can be used as a surrogate: one observes the mechanisms by which the fish develops tumours, and hypothesizes that carcinogenesis (or any subphenomenon, such as migration, angiogenesis, etc) proceeds in the same way in humans. This extrapolation, however, is warranted only insofar as the fish (and its tumour) is a replica of the human patient, in full accordance with the similarity view. As a matter of fact, however, the model was used in a quite different way, for which the tumours – and the very fact that tumours were formed – were rather superfluous.

Already at the larval stage, the transgenic fish display a hyper-pigmentation due to an

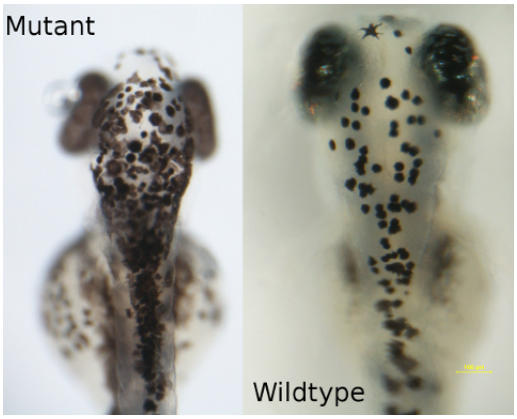

Figure 3.1: The larval phenotype of the zebrafish model.

overproliferation of melanocytes (figure 3.1). Because of the speed at which this phenotype could be observed, and of the ease with which chemicals could be administered to the animals, it could be used in a high-throughput platform to test thousands of compounds[11]. Strictly speaking, the function of this screening was not to find a drug for melanoma, but to find compounds that have *an effect on the effect of the mutant HRAS*.

This becomes obvious when we look at the operationalization of the model. The typical experiment in molecular biology has two controls: a negative control and a positive control. For instance, an experiment to measure or detect the presence of some DNA stretch in a sample will typically contain, on top of the sample to be tested, a sample that is known to contain the DNA (the positive control), and a sample that is know not to, for instance water (the negative control). Detecting the DNA in the negative control, or not detecting it in the positive control, indicates that the experiment did not proceed correctly (due, for instance, to sample contamination or bad reagents). The drug screening, however, contained an additional control. For each compound, melanocytes were counted on four groups of fish:

Wild-type untreated	Mutant untreated
Wild-type treated	Mutant treated

Table 3.1: Groups assayed in the screen.

If a compound negates the effect of the HRAS mutation, we would expect the 'mutant treated' group to depart from the 'mutant untreated' group, and recapitulate the 'wild-

[11] The information on the drug screening aspects of this zebrafish model comes from the few months I could spent under precious tutoring provided by Cristina Santoriello, at the IFOM-IEO Campus, Milano. Although the results of hte screen were not yet published, screens on very similar zebrafish models have been published, some of which are reviewed in White et al. (2013).

type untreated' group. In analogy with a basic experimental design such as the DNA detection example, the 'mutant untreated' and 'wild-type untreated' groups would represent, respectively, the negative and positive controls. However, the experimental design contained an additional control group, which had the particular purpose of controlling for effects that were opposite to the mutation and yet independent from it. For a compound could rescue the hyper-pigmentation phenotype by means completely independent from the mutation, which would most likely be irrelevant to the cancer. If a compound could rescue the phenotype, in other words bring back, in the mutant, the melanocyte count to that of the untreated wild-type, without significantly reducing the melanocyte count in the (treated) wild-type, then this compound had an effect on (the effect of) the HRAS mutation. It could then be studied further as a potential drug. The important point here is that the compounds identified are not those having an effect on the zebrafish itself, nor on its tumour: in fact we could run the same test if the mutant never developed cancer. What matters is that the compound rescues the specific mutant phenotype – whatever it is – and not simply has an inverse effect. This procedure effectively reduces the observed effect of the compound to the causal pathway connecting the mutation (the difference-maker) to the phenotype (the difference).

The genetic engineering plays here the same role that is played, in the xenograft example as well as in the case of the A-Z test, by the material transfer from target to model system. Both ensure that the difference-maker is shared between the model and the target system. Because one knows that the difference-maker (the mutation) is shared by the model and target system (as a consequence of having been engineered in that way), and because the experimental setting (with the double control) allows one to reduce the effect of a compound to a modification of the effect of the difference-maker, one is able to establish a causal interaction between the compound and the pathway through which the mutation has its effect, and to extrapolate this interaction to the target system in an imperfect but quite robust way. Compared to models as surrogates, in this context the strength of this extrapolation is considerably less dependent on the similarity between man and fish, and most importantly between human and fish tumours. In fact, in this case the tumours were not even used, and the phenotype could have been virtually anything, for it has a purely

instrumental value (see Meunier 2012). It is simply a signal for the causal relevance of the compound with respect to the immediate effects of the mutation.

The fact that this example is continuous with the xenograft cases becomes obvious when one considers more closely the genetic engineering involved. The zebrafish line was created by inserting stretches of DNA into the genome of the organism, but where did this DNA come from? The oncogene is a mutated version of the human HRAS, and at some point, hidden in the intricacies of laboratory histories, it was cloned from a human sequence extracted and isolated from tumour cells. It was then cloned through plasmids, joined to a reporter and to a promoter that would allow conditional activation, before being inserted into the zebrafish genome.

To what extent, then, is the model the result of human material being brought into the fish? If the mouse was simply an environment for the transplanted cells, so can the fish be an environment for the transplanted molecule. It is important to note that it does not make the slightest difference if the inserted DNA came directly from human cells, or if it was entirely synthetic but designed after the sequence of the human gene. What seems to matter is not that there is a material transfer, but that the mutated gene is identical to the human one: they are about as close to a natural kind as biology ever gets. We can say that the mutated HRAS is 'the same' as a human mutated HRAS in a way that is different from saying the the cells injected in the mouse are 'the same' as those in the patient's tumour, and perhaps this different warrants different kinds of extrapolation. Note, however, that the same could be said in the context of the elevated plus maze briefly discussed in section 2.4.2, where the mouse's behavior in the maze was taken as a measure of the (anti-)anxiolytic properties of drugs: from a molecular point of view, the drug used in the mouse is identical to that used in patients, and yet it seems odd to speak of the elevated plus maze as a 'molecular model in a very complex medium'. If the mouse is more than a mere environment for the drug that is administered to it – or a tool to learn about the drug – how does it differ from the transplantation examples?

3.3 Characterization of the instrumental role

Although the instrumental role of biomedical models appears intuitively very distinct from more traditional usages, the distinction is not easy to pin down in precise terms. There are a number of factors that make a given case more or less towards the instrumental role, and in this section I will attempt to make them more explicit, relying especially on philosophical discussions around the notion of instrument. Before turning to these, however, it is useful to start by relating the distinction to another proposed by Rheinberger (1997):

> "A technical product, as everybody expects, has to fulfill the purpose implemented in its construction. It is first and foremost an answering machine. In contrast, an epistemic object is first and foremost a question-generating machine." (Rheinberger 1997, p.32)

There is an important overlap between the instrumental role and Rheinberger's notion of technical object. The epistemic thing is open-ended, and in fact Rheinberger's account of the discovery of protein synthesis emphasizes the fact that experimental systems have 'a life of their own' (to paraphrase Hacking 1983). In contrast, upon using a technical object, the user knows what possible outcomes can be expected. This is indeed the case of the instrumental models presented above: although the A-Z test can perhaps dysfunction, insofar as it functions there can be no surprise as to the kind of answer it will provide. The same can be said of xenografts for the detection of cancer stem cells. However, as mentioned earlier this is not an intrinsic property of the material systems, but comes from the way in which they are used. When grafting a tumour to observe, say, the interplay of the cancer cells with their environment, the model becomes open-ended – and in what follows I will try to dissect why.

Rheinberger (1997) has discussed at length how research systems turn into technical objects:

> "If research systems become too rigid, they turn into devices for testing, into standardized kits, into procedures for making replicas. They lose their function as machines for making the future." (Rheinberger 1997, p.80)

But the transformation can also occur the other way around, and this has been much less discussed: technical objects "themselves can become epistemic things or help produce them

– if for example they elicit unexpected questions in their use." (Rheinberger 2010, p.218) This is especially clear in teratoma assays, which are used to assess the pluripotency of stem cells (discussed further in section 5.2.2; see also appendix B). The cells are injected into the mouse, and if they form a teratoma containing tissues from all three germ layers, they are considered pluripotent. Independently of whether the teratoma assay does or not display the three germ layers, if the teratoma has very particular and unexpected features, it can (and does) become the object of study. Hybrid models such as xenografts are particular in their flexibility to switch, at anytime in their use, from technical objects to epistemic things (or part thereof).

Rheinberger's distinction aligns with the replica/instrument distinction which I have tried to make in this chapter, but it does not explain it away: it merely describes a difference without explaining why there is such a difference. Nor is the alignment perfect: the mouse in the elevated plus maze is certainly a question-answering, rather than question-generating device, and as such it is more a technical object than an epistemic thing. Nonetheless, there seems to be an important epistemic difference between it and, say, the A-Z test. In this section I will attempt to clarify it by offering a more theoretical account of the instrumental role. Doing so will require some detours into different strands of thought in the philosophy of experiments and of instrumentation, which will provide some ingredients for the characterization that I will propose in section 3.3.2.

3.3.1 Production and observation instruments

The examples of the instrumental role presented earlier, especially the cases of transplantation, invite the thought that the animal is a 'mere' test-tube. This suggests that the xenografted mice discussed above (section 3.2.3) are not animal models, but rather the 'very complex medium' of an *in vitro* model. It is important to note, however, that there are very different ways in which the mouse can be said to be a mere test-tube, and this is a good place to begin the characterization of the instrumental role.

Until the 1940's, viruses were kept and grown in the lab by serial infection of host animals: in this context, the animal does not accomplish any epistemic task, but merely

serves as a practical means of keeping stocks of viruses[12]. Although the animal can be said to have an instrumental role in a very broad sense, it is not an observational instrument, let alone a measuring/detection device, for it does not serve the purpose of making anything visible. In xenograft experiments meant to detect the presence of CSCs, however, the mouse does have such a purpose, and hence it can be said to be an instrument in a much narrower way. Importantly, the same can sometimes be said with respect to *in vitro* models: while the medium in the dish is often designed for the mere practical aim of maximizing cell growth and survival, it also often has the purpose of making certain things visible (e.g. live stainings, dye dilution assays, tagged nucleotides, etc.).

This distinction is an important one, applying equally to cells *in vitro*. It was problematized early on in the application of tissue culture techniques for drug screening[13] and is still relevant today. For instance, pluripotent cells are traditionally grown on a layer of feeder cells (inactivated mouse embryonic fibroblasts) which produce necessary signals for the survival and maintenance of the pluripotent cells. Although these culture conditions bear some resemblance to physiological stem cell maintenance (in which the niche is particularly important), they became a standard because they allowed the efficient growth of cells and the maintenance of their pluripotency. It could be replaced with anything that would allow the maintenance of these cells, independently of whether this models the tissue environment or not, and in fact they are being replaced by feeder-free conditions, not for the sake of a greater similarity with the cells in a body, but for practical reasons: most importantly the difficulty, in subsequent experiments (for instance in Next Generation Sequencing experiments), to separate signals from the feeder from those from the pluripotent cells.

A first attempt to cash out these different kinds of instrumental roles would be to say that the role of some instruments – medium, animal, etc. – is epistemic, while others have a practical role. Of course, instruments with a 'merely' practical role have this role

[12] Gradmann (2006) reports that Ferdinand Hueppe, already in 1885, was talking of animals as "milieu de culture vivant" – living media for the culture of germs. In the middle of the 20th century, especially following the work of Enders, cell culture was instead used as a mean of keeping stocks of viruses.

[13] For instance, in a workshop on the topic, a participant (L. Berman) remarked that: "Much of the recent work with cultured cells is based on techniques lately developed by virologists. These have been designed to promote maximum proliferation of cells. [...] It is doubtful whether we can apply all of our past experience with tissue culture to the present type of cell culture, which has the practical aim of obtaining large quantities of living cells, rather than the study of the cells themselves." (Eagle 1958, p.549-550)

within an epistemic endeavor, and instruments with an epistemic role have a major practical dimension in that they always produce something. In order to clarify the distinction, it is useful to rely on some concepts proposed in the philosophy of experimentation.

Michael Heidelberger (2003) identifies different kinds of experiments in which scientific instruments are involved (each of which, he argues, is affected by different kinds of theory-ladenness). He first identifies two kinds of experimental manipulations involving instruments, which he calls causal manipulations and representations. The first involves the production of phenomena, while the second involve observational 'readouts', most explicit in measurement devices.

The first form, he argues, comes into three kinds:

1. Productive role: "to produce phenomena that normally do not appear in the realm of human experience." (Heidelberger 2003, p.146)

2. Constructive role: "to produce an effect in its 'pure form,' without any complications or additions that could spoil it or that are otherwise alien to it. Another goal is to tame the phenomena in order to be able to manipulate them in a certain desired way." (Heidelberger 2003, p.147)

3. Imitative role: "to produce effects in the same way as they appear in nature without human intervention." (Heidelberger 2003, p.147)

In biology, it seems at first unclear to what extent the three kinds are distinguished. The productive role seems to be related to whether or not the phenomena normally appear in nature or human experience, but this obviously depends on the definition of the phenomena. Consider, for instance, the zebrafish model discussed earlier (section 3.2.4). Allegedly, there are such things as 'HRAS-mutated tumours of melanocytic origin' happening naturally in humans. Specifically, however, the phenomenon presented by the model is rather a tumour developing in a fish melanocyte because of mutated human HRAS expressed under a GAL4-UAS system – something clearly alien. The trouble is that biological phenomena, unlike perhaps some phenomena of fundamental physics (although Hacking 1983 has convincingly argued that they are equally constructed), are not given already partitioned, and there does not seem to be an absolute way of choosing between different partitionings. This

being said, some biological experiments are explicitly aimed at producing entirely artificial phenomena, or at least phenomena whose natural occurrence is irrelevant. For example, in most cases it does not seem to make a difference to the meaning of a knock-out experiment whether a similar mutant mouse was ever born naturally. Instead, the experiment produces a phenomenon (a mutant phenotype) whose occurrence in human experience is irrelevant because the phenomenon is but a means of investigating something else, namely the knocked-out gene's function in normal (i.e. wild-type) physiology. However, when the knocked-out gene is known to be also inactivated in a monogenic disease, the situation seems different: suddenly, the mouse's phenotype is perceived in an analogy to the disease – it is imitating rather than producing. Obviously, there is also an important constructive role: human patients affected by the disease each have a host of other genetic variations, while the knock-out experiment produces the diseased condition in its 'pure' form, in the near-absence of background variations; furthermore, mice provide the opportunity for countless manipulations and assays that are unavailable with patients. All experiments, at least in biology, involve a component of Heidelberger's constructive role. Nevertheless, the mouse is imitating in a way that other knock-out models are not. Why does it make a difference to the meaning of the knock-out experiment whether the gene is involved in a monogenic disease or is largely unknown? The difference, it seems, is not in the production of the phenomena, but in the way we look at it: if the mouse is imitating a monogenic disease, we will look at its phenotype within the reference frame provided by the disease. For example, if the disease has a cognitive component, we will concentrate on behavioural assays or neurophysiology. In other words, *the different kinds of producing are relevant insofar as they suggest different kinds of looking at the products*.

Most observations are acts of production (Hacking 1983, Rheinberger 1997), and production experiments have an outcome only insofar as it involves, or is coupled with, observation. This threatens to undermine a distinction between Heidelberger's two forms of experimentation (and similar distinctions, such as that proposed by Baird (2003)). I believe however that there is still a relevant distinction to be made, for *there are modes of producing that are closely tied to specific modes of representing*, while there are others that are more loosely related to a broad set of modes of observing. A thermometer, for instance,

produces a phenomenon (the rise and fall of a mercury column), but this phenomenon is useful only as a measurement of temperature. In contrast, a mouse knock-out experiment is open to virtually any number of relevant kinds of observations. Experimentation of the first form differs in that it is largely decoupled from, or put very little constrain on, the kind of observation one can make about the product.

Heidelberger describes his second form of experimentation "as adjustment to a theoretical context or assimilation to a theoretical interpretation with the help of instruments", in which

> "the goal is to represent symbolically in an instrument the relations between natural phenomena and thus to better understand how phenomena are ordered and related to each other. [...] These are 'information-transforming instruments,' as Davis Baird once called them; they transform the input information into a more useful output format while preserving the order of the phenomena vis-à-vis the intensity of the attribute in question." (Heidelberger 2003, p.147).

This second form is typical of measuring experiments. As van Fraassen puts it, "measuring locates the target in a theoretically constructed logical space." (van Fraassen 2008, p.2). A measurement is a projection of some phenomena into a constrained space of representation. A thermometer, for instance, is used to locate, say, a body of liquid, with respect to (states of) other substances. Early thermometry ordered substances on an arbitrary scale built between fixed points, which were established and unsettled throughout the history of the discipline: the freezing point of water, the melting point of ice, the boiling point of water, the melting point of butter, blood temperature, the first night frost, etc., up to the temperature of the cellars of Paris' Observatory (see Chang 2004, p.10). Contemporary thermometry locates substances with respect to theoretically fixed points (the absolute zero kelvin) or empirical fixed points (e.g. degrees centigrad), using an arbitrary (but theoretically defined) scale. In any case, the thermometer is used to locate a substance on a space of representation – a map.

Every observation is theory-laden (in the sense described in section 1.1.3), and as such every observation experiment locates the phenomena with respect to *something else*. When we observe, for instance, that a mutant mouse has a given phenotype, this phenotype is defined with respect to the normal (wild-type) phenotype, and to the kinds of differences we have learned to pay attention to. But do all these reference points deserve the label of

'map', or "theoretically constructed logical space" (van Fraassen 2008, p.2)?

3.3.2 Replica and instruments

The point is important, because unless we further restrict the instrumental role, the vast majority of examples of animal models (or, for that matter, any experimental set-up) could be represented in this way. Consider, for instance, the traditional case of an animal being used as surrogate to test the carcinogenicity of a chemical compound, say tar. Tar is regularly applied (the input) on mice, and after some time the incidence of cancer (the signal) is measured. If there is a significantly higher incidence of cancer in the treatment versus the control group, then this is evidence for the carcinogenicity of tar. Hence the signal is used to learn something about the input – tar. While I would argue that there is a value in viewing every form of experimentation in this angle, it is useful for the purpose of illustration to characterize a narrower instrumental role. Such a role can then be used to emphasize important differences in the way biomedical models are used.

The simple tar-test relies on carcinogenicity in the mouse to extrapolate carcinogenicity in human, and therefore (by assumption) informs us of the carcinogenicity of the substance on both organisms at the same time. This is the first assumption – surrogacy – discussed in the previous chapter (section 2.4.1). This symmetry is an important property of surrogacy – as the NRC committee puts it, "the modelling relationships are reciprocal" (Committee on Models for Biomedical Research 1985, p.19). Indeed, one could equally use, in the same experimental set-up, carcinogenicity in humans as a signal for carcinogenicity in the mouse. Measuring devices, on the other hand, typically do not share this feature. This is because such instruments involve a decoupling between the output of the instrument – the signal – and the information this signal provides on the target system (from which the input comes). In fact, in many cases the signal observed in the device has no clear counterpart in the other system (there is no meaningful way to say of a pregnancy test bought at the drugstore that it is pregnant, or to say of a woman that she has 'two colored bands').

In the Ascheim-Zondek test (section 3.2.1)), for example, the 'Blutpunkte' are not meant to infer the presence of blood stains in the woman, but rather of pregnancy (and the mice are clearly not pregnant). In this context, the extrapolation is not essentially

based on similarity, or at least not from the point of view of justification, for what is being extrapolated is very dissimilar. Instead, the inference is based on simple induction and calibration – a process by which the instrument is modified in order to provide a better signal.

The biologist, during the dissection of the mouse, may notice a variety of phenomena: perhaps the mouse's ovaries have a peculiar shape, a greenish taint, or the mouse has a particularly pungent smell. Or perhaps the follicles have particularly many of these small blood stains – but there is no way in which the woman is 'more pregnant', and from these facts the biologist will never draw any conclusion *for the purpose of the pregnancy test*.

Measuring devices locate their object on a theoretically constructed space, and the notion of space suggests two important things. First, a space (at least in a Cartesian outlook) is a structure that is to some extent defined beyond the objects that populate it: an area in a space is defined and meaningful even if nothing occupies it. Second, a space has defined *dimensions*. Consider the example of early (xeno)transplantation as a measurement procedure (section 3.2.3). The procedure is tied to the creation of a logical space – initially binary and then numerical – in which tumours were located. To some extent, this space was structured by the transplantation system: tumours engrafted or not, or engrafted a certain proportion of the time. But the fact that it was unidimensional required a step of abstraction. Measuring devices imply a double restriction of dimensionality: only part of what is observed in the apparatus is actually informative about only some aspects of the input. They are projections analogous to the topological sense: much like the projection of the earth on a two-dimensional plane, they combine and reduce the dimensions of what they represent (they are not bijective). And as we know from topological projections, there is no perfect projection, but only projections that are more useful for given purposes.

Many measuring devices constrain their object to an even greater extent, and restrict the possible readouts within a single dimension. Here it is useful to rely on the analog/digital distinction (Goodman 1968), which was recently applied to traits/phenotypes by Meunier (2011). The following example, while simplistic with respect to Goodman and Meunier's respective discussions, is sufficient for our present purposes. A digital watch, for instance, tells us the time in a single and definite way: there are not different ways of reading the

watch to learn about the time, and one will not get more information by looking at it more closely. In other words, it is straightforward to say whether two watches are giving the same time. A gauge, by contrast, is analog (between any two points on the gauge is always another one), and in practice there is no saying that two measurements are the same (would they still be the same under a magnifier?). Importantly, digital does not imply quantitative, and in fact Goodman discussed it primarily for non-quantitative settings. What it does however imply is 'notational', which makes the readout linked to a representational system. Indeed, being digital or analog is a property not of an object, but of the whole system of representation (a digital watch may be used as an analog for other purposes than knowing the time).

Reduction of dimensionality and digitalization have an important epistemic function. Because anything is in some way similar to anything, and anything is in some way different from anything but itself, there are countless way of grouping and discriminating things. Constrains placed on the space of representation gradually reduce this difficulty: in the simplest case, a binary space, elements are in one of two groups.

We can therefore characterize the two usages of biomedical models. The first is the more traditional animal model as conceived by the surrogate view, the paradigmatic example of which is the dissection of an animal, observation of a phenomenon, and inference of the same phenomenon in humans on the ground of phylogenetic proximity (as a proxy for similarity). I will refer to this role as the surrogate role, which requires the model to be at least a partial replica of the target system. On the other hand, the animal can be used as an instrument, and more particularly as a measuring or detection device. This involves a decoupling between what is observed in the model (the signal) and what it allows to infer about the target, which is tied to the fact that the measuring device provides a projection of the phenomena in a constrained space of representation. This generally implies both a reduction (and transformation) of dimensionality of the phenomena, and some extent of digitalization.[14]

[14] The surrogate role shares important similarities with what Gayon (2006) referred to as the 'exemplar' aspect of model organisms. However, Gayon's notion of exemplarity has an additional component, namely the idea of the organism being representative of a broad (and necessarily varied) class of organisms – the model organism is in a way an 'ideal organism'. Surrogacy, on the other hand, replaces an organism with another, and therefore does not imply such a broad target class. Likewise, the instrumental role proposed here is in some ways similar to what Gayon (2006) labeled as 'tool' ("organisme-outil"), although his term is more encompassing than the role I described.

This being said, these two roles likely represent the two ends of a continuum. I would argue that every model is a projection on a different space of representation, although some such spaces are too vaguely defined for such a view to be useful.

3.3.3 The growing importance of the instrumental role

The example of the engineered zebrafish (section 3.2.4) shows that the instrumental role is not limited to some anomalous or borderline cases like xenotransplantation, but is also very likely to be common in many other fields, especially (but not exclusively) where genetically engineered models are used. In fact, I would claim that this use is becoming more and more important, because of related reasons which are worth briefly mentioning.

The first reason has to do with reductionism: the zebrafish example yields information that can mostly be rendered at the macromolecular level. In such a context, organisms are but a way to probe molecules. This general strategy does not need to assume that phenotypes are reducible to the molecular, but simply that relevant knowledge – in the case presented, means of intervention – can be gained at the molecular level. It also assumes a relative generality of molecular interactions (if a compound A interacts with a pathway B in a given context, then it does so in most contexts), or at least a comparability of context between the systems. This assumption is pervasive in contemporary biology, especially molecular biology[15], which does not mean that it goes untested, but simply that it represents a sort of default hypothesis (and productively so).

The second reason is the scientists' increasing capacity to craft their models (Maugeri and Blasimme 2011). As I showed with the zebrafish example, genetic engineering can also fulfill the same role as the material input from the target system: that of ensuring that the difference in which we are interested is the consequence of a same difference-maker. Obviously, this in the first place requires having an actual difference-maker in the target system. This is only possible if the target system is narrow and homogeneous enough. As opposed to biology in general, biomedical research is interested in gaining knowledge about humans and human pathologies, and this narrow target class enables the crafting of animal

[15] It is for instance telling that interaction databases generally do not include standardized fields to specify the cell type in which the interaction was observed. See for instance the PSI-MI format for storage of molecular interactions, used notably by the BioGrid.

models specifically geared toward extrapolation.

Finally, the fourth reason is the distributed nature of biomedical research. As mentioned in the previous chapter, much of the discussion on biomedical models assumes a simplistic perception of biomedical research inspired by massive screenings such as the early CCNSC. This is often reconstructed as a linear and unidirectional process in which each model is taken to be independent from the others (see sections 2.4.2 and 2.4.3). As I have argued in the previous chapter, this view of drug discovery is unrepresentative of contemporary research, and is even inappropriate to understand the last three decades of the CCNSC itself. Instead, I will argue in Chapter 5 that biomedical research is best understood as distributed modeling: a highly non-linear way, with scientists shifting between different models, transferring samples and materials from one to the other, introducing insights from the clinic into their bench work, and so on. In this more elaborate way of knowing, it is not anymore a single biomedical model that is expected to predict clinical outcomes, but a network of interacting model systems. As a consequence, each such system is not expected to act as a surrogate for a human patient, and its scope is much narrower: like other instruments, they are small nodes of a research system, each meant to answer highly specific questions. The reductionism mentioned earlier is not so much the reflection of a reductionistic conception of disease, but rather of this changing way of doing biomedical research.

3.4 Proximate functions of biomedical models

A central claim of this chapter was that biomedical models inform us about human pathologies in a variety of ways which are not exhausted by traditional accounts of animal models. The distinction between models as instruments and models as surrogates is one of the most important for epistemological analysis, but it is nonetheless only one of many relevant distinctions that could be made. It is therefore useful to sketch a taxonomy of the proximate functions of biomedical models. That is to say, a distinction not on the basis of structural features of the models in themselves, nor of the ultimate goals they are meant to further, but rather on the way these models connect with the rest of the research system in which they are embedded.

3.4.1 Models as tools for thinking

Some models are used to represent or exemplify a prototypical mechanism in very much the same way one would do by sketching it on a piece of paper, although with biological means. Weber (forthcoming) writes of 'in-vivo representations' of natural processes, and characterizes this practice as *experimental modelling*. Weber for instance claims that experimental evolution (Lenski and Travisano 1994, Ratcliff et al. 2012) is an instance of experimental modeling because it aims at representing, through living means, a general and idealized (often simplified) mechanism rather than any specific one. This is very much like a proof-of-concept of a mechanism (or proof-of-principle), which has been important throughout the history of molecular biology and has become even more common in synthetic biology, where abstractly constructed mechanisms gain credibility by finding a biological implementation. But it is also very similar to computer models in that it provides an understanding of the dynamics of a mechanism: it can for instance "show under what parameter values a certain phenomenon is possible" (Weber 2012). Here, the representational scope of the model is not defined in terms of biological classes, and the model is rather a possible interpretation of a theoretical construct. Like scientific models more generally, experimental models are "tools for thinking" (Knuuttila 2011), or even a form of theorizing (Waters 2012).

3.4.2 Models as replica or surrogates

These are the models as understood by the surrogacy view (section 2.4.1) : the relation of representation holds between the model and a target system or class. The relation is symmetrical and reciprocal (model and target could be swapped without major change to the experimental design), and extrapolation is performed through analogical reasoning, which implies that it is warranted almost exclusively on the basis of similarity between the two systems, although the similarity can be only with respect to a relevant subset of features. Although it does not mean that extrapolation cannot be controlled (see for instance Maugeri and Blasimme 2011), this usage of biomedical models is subject to the criticisms raised against the predictive use of animal experimentation (section 2.2.1). However, as discussed in Chapter 2, even though animals are poor surrogates for humans, they can nonetheless be useful surrogates, especially in a context in which they are but one step in a more complex

screening apparatus.

3.4.3 Models as instruments

This is the model as instrument or measuring device characterized earlier (see section 3.3, especially section 3.3.2). Organisms can have many instrumental uses, some of which are better kept distinct. I will restrict the term instrument to the models used as observational instruments – as means of making otherwise invisible differences visible involving a reduction in dimensionality and/or some form of digitalization. Detection devices are those where the signal is binary (e.g. yes/no, pregnant or not, etc.), whereas other measuring devices allow for gradation or quantification (e.g. degrees of transplantability, etc.). In all cases, the model produces a phenomena that is closely tied to a specific – and, most often, theoretically embedded – mode of observation or recording.

3.4.4 Models as reagents or factories

Cells and organisms are sometimes used in a purely technical way, and have no specific epistemic import. They are purely material means of obtaining or doing something material, in an exchangeable way. Early pathologists were using animals to grow and maintain stocks of germs, in which case the only use of the organisms is to keep the viruses and micro-organisms alive while no work is done on them. Likewise, tumour transplantation also started as a means of growing and maintaining tumours. Nowadays, plasmids and bacteria are routinely used in genetic engineering, and so are rabbits, mice and cells for the mass production of antibodies. The antibodies produced have nothing to do with human antibodies (in fact they generally target human proteins), and yet they are essential to biomedical research (see Cambrosio and Keating 1995). All of these examples fit into what Shanks and Greek called 'bioreactors' or organisms as 'factories' (Shanks and Greek 2009, p.30). However, as discussed earlier, some means of producing material things are different from others in that they are tied to modes of observing – these belong to the previous category (models as instruments). In contrast, models as reagents or factories are exchangeable, and not tied to any mode of observing. This is not to say that means of production are epistemically irrelevant: for instance, monoclonal antibodies are known to

be more precise in they affinity than polyclonal antibodies. However, the fact that they are of one type or another, or the fact that they were made in a goat instead of a rabbit, is largely irrelevant to the way they are then used in the lab[16]. The difference in quality is something that can be judged by assays on the antibody batch itself, irrelevant of its source.

3.5 Conclusion

As different authors have noted, "model organisms have epistemic functions over and above providing a basis for inductive inferences or extrapolations to other organisms" (Weber 2005, p.182). Understanding these functions is necessary to an epistemology of disease modelling, and to any attempt at evaluating disease models. In this chapter, I have tried to explore the diversity of proximal functions fulfilled by biomedical models, concentrating especially on models as observational instruments, which are somewhere in between mere reagents, and the traditional model as surrogates. Such biomedical models have the purpose of making otherwise invisible differences visible, and they do so by locating an entity or phenomena on a map, or theoretically constructed space of representation.

For this reason, measuring devices are informative in pre-determined or constrained ways, which offers opportunities for theorizing, but can become a limitation. Among the reasons explaining the dramatic drop in efficiency of pharmaceutical research, Scannell et al. raise an issue which they call 'the narrow clinical search problem': "the shift from an approach that looked broadly for therapeutic potential in biologically active agents to one that seeks precise effects from molecules designed with a single drug target in mind." (Scannell et al. 2012, p.197) In the 1950's and 1960's, they argue,

> "[d]iscovery involved, to an extent, the ability of physicians to spot patterns through careful clinical observation, especially in therapeutic areas in which symptomatic improvements are readily observable, such as psychiatry."

Instead, nowadays

> "[i]f a drug has an effect but this is not the precise effect that the trial

[16] Goat and rabbit antibodies obviously require different secondary antibodies, but this does not put much restriction on the way the are used.

> designers anticipated, then the trial fails. Opportunities for serendipity are actively engineered out of the system." (Scannell et al. 2012, p.197)

At the core of the problem is, of course, a trade-off between throughput and flexibility. Although much broader, this problem seems exacerbated in the instrumental use of biomedical models. In many cases, the artificiality of biomedical models used as instruments prevents them from becoming relevant replica: for instance, the mouse of the A-Z test (section 3.2.1) is a poor replica for a pregnant woman, and as such its utility seems restricted to its instrumental role. Other cases, such as xenografts, are different in that they can change, in the course of their use, from instruments to open-ended epistemic things (section 3.3). This flexibility is arguably an added value of such systems.

I have argued that the instrumental role is becoming more prominent, and I want to argue that there is a value in considering biomedical models in general as projections in different spaces of representation. Since these spaces are theoretically constructed – more or less precisely depending on the case at hand – we must consider the theoretical terms and frameworks involved in biomedical modeling. The relationship between biomedical models and the theoretical language will be the topic of the next chapter.

Chapter 4

Models and theory

4.1 Introduction

The previous chapter described a prominent and yet under-appreciated usage of biomedical models which I labeled the instrumental role. Following van Fraassen's treatment of measuring devices (van Fraassen 2008), I argued that these models are characterized by the inscription of their readouts into a 'map', or 'theoretically constructed space'. It is therefore important to understand the relationship between biomedical models and theoretical terms or frameworks, and this will be the aim of this chapter.

Theory was traditionally conceived in philosophy of science as an abstract and universal syntactic structure, a notion which has been recognized as ill-suited for biology (see for instance Keller 2000, 2002). This is not to say, however, that biology has no theoretical discourse, and a generalized disdain for theory might be as problematic for the history and philosophy of science as an obsessive focus on theories. Weber has made this point several times in his discussion of 'New Experimentalism', for instance:

> "New Experimentalism arose as a movement to counter the extreme theory-centrism that has dominated classical philosophy of science, where experiments entered at best in the form of an 'observation sentence *o*' or 'evidence *e*'. But as usual, welcome attempts to redress the balance sent the pendulum to the other extreme, leading to a discourse about science that completely ignores theory. It is time to emphasize that theory and experiment are equally important and deserving of philosophical and historical scrutiny. What New Experimentalists have tended to overlook is that experimental systems always come with *theoretical interpretations* of what happens in an experiment." (Weber 2011, p.218)

As the notion of theory[1] is notoriously difficult to define, my working definition for a theoretical term will be a term which does not have an exhaustive operational definition – in other words, that either has no operational definition, or that can have different conflicting ones. Theoretical structures and relations are those that involve theoretical terms.

Drawing such a distinction suggests that there are, in some strong foundational sense, terms that have an exhaustive operational definition. I do not believe that this is the case, for the simple reason that it leads to an epistemic regress in which one should acknowledge having applied the right operations. My point is about representations more generally. Nevertheless, some terms are farther from operations than others, and all I want to do with the notion of theoretical is to concentrate on those cases where operations are less tightly defined, and which will consequently make more obvious the issues I want to discuss.

The first part of the chapter (section 4.2) discusses the role of the theoretical realm in extrapolation, relying on two historically important attempts at a philosophy of science. Section 4.2.1 discusses Claude Bernard's approach to extrapolation from animal studies, and section 4.2.2 discusses the epistemological solution proposed to an analogous problem in psychology. I want to argue that theoretical concepts often work as bridges between material systems, but that these bridges cannot be conceptualized as simple correspondences. Furthermore, I argue that these bridges ground not only animal experimentation, but simultaneously medicine as a whole.

In the second part (section 4.3), I show how these discussions relate to examples of xenograft experiments in cancer research (which were already introduced in the previous chapter, section 3.2.3). I discuss the relationship between instrumental biomedical models and their related theoretical terms. Using the example of the recent debate on melanoma-initiating cells (section 4.3.4), I argue that the meaning of the theoretical notion of 'cancer stem cells' cannot be reduced to operations, and conversely that the appropriate xenograft model for the detection and study of this theoretical construct cannot be determined independently of the theoretical relationships the term entertains. The upshot of this discussion is that at least for the instrumental role, the evaluation of biomedical models cannot rest on a similarity between the model and a clinical counterpart, but instead depends on the

[1] Some colleagues and I have argued elsewhere (Blasimme et al. 2013) that an adequate account of scientific developments in biology requires the recognition of some kind of theoretical apparatus or higher-level conceptual framework – what we called 'explanatory frameworks'.

theoretical framework in which it is embedded. Finally, in section 4.4 I related these points to a broader, coherentist view of science.

4.2 Extrapolation and the theoretical

Several authors have pointed out that the use of animals in biomedical research, and in fact the very idea of 'model organisms', presupposes the existence of 'generic relationships' (Creager et al. 2007, p.2) – of a 'General biology' (Rheinberger 2006, p.46; Rheinberger 2010, p.6) that would apply across species[2]. Indeed, the expression 'General biology' (capitalized) was explicitly used in this sense by the NRC's Committee on Models for Biomedical Research (Committee on Models for Biomedical Research 1985, p.10), and as I will discuss in the next section, Claude Bernard's method and philosophy were also founded on the same logic, although he instead wrote of a 'theoretical medicine'. In all these cases, however, it is important to note that adopting a notion of 'General biology' does not presuppose that everything observed in an organism will be generalizable to others, but rather that there are kinds of knowledge which will be generalizable. Therefore, one of the main claim of this chapter will be that this approach requires a redefinition of the scientific object with extrapolation in mind.

4.2.1 Claude Bernard and the foundations of experimental medicine

Although Claude Bernard (1813-1878) is generally considered the father of modern medicine, it has been argued that he asserted the primacy of laboratory science, especially animal experimentation, at the expense of a clinical medicine which he allegedly 'denigrated' (see for instance LaFollette and Shanks 1994, p.196). To be precise, Bernard asserted the importance of experimentation, but did not doubt the importance of clinical observations (and in fact argued that it was a necessary component of medicine – see Bernard 1865, third part, chapter IV). He did however repeatedly emphasize that medical science could not stop there – "la médecine ne doit pas en rester là" (Bernard 1865, third part, chapter IV

[2] Rheinberger (2006) cites Hartmann's *Allgemeine Biologie* (1927) as an example of a 'General biology', but there are many examples of this endeavour: intermediary metabolism, population genetics, allometric laws, or principles such as homeostasis. The notion of 'biological problem' in (Hubbard 2007, p.60), as well as Schaffner's 'middle-range theories' of restricted scope (Schaffner 1993, p.97-98), both suggest some sort of local generality.

§IV). What he denigrated was a clinical practice that did not attempt to go from symptoms to physiological causes, and tended towards a kind of indeterministic nominalism. To Bernard, animal experimentation was to some extent a practical necessity, due to the major restrictions of the application of the experimental method to human beings. This narrative however obscures the fact that Bernard's philosophy of science was not simply an ad hoc justification of extrapolation from animal experiments, but was the very condition of possibility of any medical science.

To see why it is so, it is useful to return to some earlier remarks. In Chapter 1 (section 1.2.1), I argued that extrapolation between members of a same species, just like cross-species extrapolation, is based on an imperfect similarity, generally inferred from genetic proximity. In Chapter 2 (section 2.2.1), I remarked that as a consequence, the arguments of Shanks and Greek (2009) against the value of animal experimentation applied equally well to clinical trials. This is indeed a critical issue: for many major diseases, drugs have therapeutic efficacy in about 50%-70% of patients (Spear et al. 2001), often with side-effects. Drug adverse effects are estimated to affect millions of persons and incur major expenditures every year. But in Bernard's time, inter-patient variability threatened the very scientificity of medicine. Bernard reports a criticism addressed to him which nicely illustrates this:

> "You say that in physiology the results of experiments are identical when they are performed under identical conditions; I deny that it is so. [...] Every time life is involved in phenomena, he added, even though it might be in identical conditions, the results can be different." As a proof of his opinion, Gerdy quoted cases of individuals affected by the same disease and to whom he had given the same drugs, and whose outcome had been different. [...] Everyone remarked to Gerdy that his opinions were nothing less than the negation of the biological science and that the identity of conditions in the cases he mentioned was illusory, in the sense that the diseases he saw as identical were not, and that he attributed to the influence of life should instead be attributed to our ignorance in phenomena as complex as those of pathology.[3]

[3] " 'En effet, vous dites qu'en physiology les résultats des expériences sont identiques quand on opère dans des conditions identiques ; je nie qu'il en soit ainsi. [...] Toutes les fois, ajouta-t-il, que la vie intervient dans les phénomènes, on a beau être dans des conditions identiques, les résultats peuvent être différents.' Comme preuve de son opinion, Gerdy cita des cas d'individus atteints de la même maladie auxquels il avait administré les mêmes médicaments et chez lesquels les résultats avaient été différents. [...] Tout le monde fit remarquer à Gerdy que ses opinions n'étaient rien de moins que la négation de la science biologique et qu'il se faisait complètement illusion sur l'identité des conditions dans les cas dont il parlait, en ce sens que les maladies qu'il regardait comme semblables et identiques ne l'étaient pas du tout, et qu'il rapportait à l'influence de la vie ce qui devait être mis sur le compte de notre ignorance dans des phénomènes aussi complexes que ceux de la pathologie." (Bernard 1865, third part, chapter II §IV)

Bernard's philosophy, therefore, was simultaneously safeguarding the scientificity of medicine and legitimizing the use of animal experimentation to learn about human biology. And it is clear that for him, the two go together:

> The different animal species offer numerous and important differences in pathological tendencies. [...] The experimental study of these diversities can furnish an explanation of the individual differences observed in man, either in different races or in different individuals of the same race, differences which physicians call predispositions or idiosyncrasies.[4]

To present Bernard's foundational thinking, it is useful to start with a famous passage which is often taken to exemplify a fundamental principle of toxicology:

> It has been observed by able and accurate experimenters that the venom of a toad quickly poisons frogs and other animals, while it has no effect on the toad itself. [...] Here again is a crude fact which could become scientific only on condition of knowing how the venom acts on a frog, and why it does not act on the toad. This necessitated the study of the mechanism of the death, for particular circumstances might be encountered which would explain the difference in outcome between the frog and the toad. Thus a special arrangement of the nostrils and the epiglottis explains very well why, for example, section of the two facial nerves is mortal in horses and not in other animals. But this exceptional fact is nonetheless rational; it confirms the rule, as we say, in that it makes no fundamental change in the nervous paralysis which is the same in all animals. There was nothing of the kind in the case with which we are concerned: study of the mechanism of death by toad's venom led to the conclusion that toad's venom kills by stopping the heart in frogs, while it does not act on a toad's heart. [...] I decided to repeat the experiments, even though I did not doubt their accuracy as crude fact. I then saw that toad's venom easily kills frogs with a dose that is wholly insufficient for a toad, but that the latter is nevertheless poisoned if we increase the dose enough. So that the difference described was reduced to a question of quantity and did not have the contradictory meaning that might be ascribed to it.[5]

[4] "Les diverses espèces d'animaux nous offrent des différences d'aptitude pathologiques très nombreuses et très importantes ; [...] Or, l'étude expérimentale de ces diversités peut selon nous donner l'explication des différences individuelles que l'on observe chez l'homme, soit dans les différentes races, soit chez les individus d'une même race, et que les médecins appellent des prédispositions ou des *idiosyncrasies*." (Bernard 1865, second part, chapter II §VII)

[5] "Il a été vu par des expérimentateurs habiles et exactes que le venin du crapaud empoisonne très rapidement les grenouilles et d'autres animaux, tandis qu'il n'a aucun effet sur le crapaud lui-même. [...] C'est là encore un fait brut qui ne pouvait devenir scientifique qu'à la condition de savoir comment ce venin agit sur la grenouille et pourquoi ce venin n'agit pas sur le crapaud. Il fallait nécessairement pour cela étudier le mécanisme de la mort, car il aurait pu se rencontrer des circonstances particulières qui eussent expliqué la différence des résultats sur la grenouille et sur le crapaud. C'est ainsi qu'il y a une disposition particulière des naseaux et de l'épiglotte qui explique très bien par exemple pourquoi la section des deux faciaux est mortelle chez le cheval et ne l'est pas chez les autres animaux. Mais ce fait exceptionnel reste néanmoins rationnel ; il confirme la règle, comme on dit, en ce qu'il ne change rien au fond de la paralysie nerveuse qui est identique chez tous les animaux. Il n'en fut pas ainsi pour le cas qui nous occupe ; l'étude du mécanisme de la mort par le venin de crapaud amena à cette conclusion que le venin de crapaud

Shanks and Greek (2009) (following LaFollette and Shanks 1994) interpret this passage in the following way:

> "Bernard here lays down one of the basic principles of the science of toxicology: once purely quantitative differences have been allowed for (differences in metabolic rate, body weight, surface area, etc), same cause will be followed by same effect in members of a given species, or in members of different species." (Shanks and Greek 2009, p.68)

In toxicology, scaling rules are used to convert, for instance, doses tolerated in animals to human-equivalent doses (see for instance Gad 2007, p.843). A similar interesting example is the reasoning of Hermann J. Muller (1890-1967) when he was asked to give his opinion to the US BEAR Committee (Biological Effects of Atomic Radiation, 1954-1964), which had the complex task of assessing the potential effects of radiation on human health. The effect (mutation rate) for different exposures was already known in *Drosophila* from Muller's work, and therefore Muller recommended the same research in inbred mice on the ground that by seeing the difference in effect between the two species, one could then approximate the effect in man by simply applying the same scale (Rader 2004, p.248). As required by the circumstances, the strategy was precautionary: the idea was not that the mouse is 'halfway' between flies and humans, but instead that the change in effect from mouse to man *could not be more dramatic than that from fly to mouse*.

Shanks and Greek (2009) are very critical of scaling rules in toxicology, arguing that different compounds scale differently, and that some species differences are qualitative. In fact, Bernard himself was already suspicious of simple scaling rules[6], and the differences he sought were not always quantitative (consider the example of the horse above). Nevertheless, the logic described by Shanks and Greek seems a broadly adequate description of Bernard's thinking. Bernard argued that as in other sciences, there are laws governing biological phenomena – sometimes called 'lois vitales' or 'lois naturelles' – and that they are what the scientist ought to seek. Phenomena are but the manifestation of these laws in

tue en arrêtant le coeur des grenouilles, tandis qu'il n'agit pas sur le coeur du crapaud. [...] j'ai voulu répéter des expériences, bien que je ne doutasse pas de leur exactitude, comme fait brut. J'ai vu alors que le venin de crapaud tue la grenouille très facilement avec une dose qui est de beaucoup insuffisante pour le crapaud, mais que celui-ci s'empoisonne néanmoins si l'on augmente assez la dose. De sorte que la différence signalée se réduisait à une question de quantité et n'avait plus la signification contradictoire qu'on pouvait lui donner." (Bernard 1865, third part, chapter II §II)

[6] "Je signalerai encore comme entachée de nombreuses causes d'erreurs la réduction des phénomènes physiologiques au kilo d'animal." (Bernard 1865, second part, chapter II §IX)

particular conditions (see for instance Bernard 1865, first part, chapter I §IV). Because laws are distinct from the conditions in which they operate, it is possible to make general claims about the venom despite idiosyncrasies in its action. This is very similar to traditional discourses in other sciences, as shown by the following analogy:

> "The following analogy – of which Bernard himself was probably aware – may illuminate his proposal for dealing with idiosyncrasies. In 1846, the French astronomer Leverrier, aware of some seemingly anomalous and idiosyncratic motions of the planet Uranus, predicted the existence of a hitherto unobserved planet Neptune. Neptune was subsequently found within a degree of its predicted location, and order was once again restored to the Laplacian universe.
>
> Whether he was aware of this particular case or not, Bernard evidently thought we could find physiological 'Neptunes' which would explain observed idiosyncracies. Differences were to be explained by more fundamental similarities; they were not physiologically irreducible." (LaFollette and Shanks 1994, p.202-203)

In other words, the meaning and utility of laws – for instance of scaling rules in the context of toxicology – depend on a proper characterization of the conditions in which they are to act. It would be a mistake to reduce toxicology to a simple rule of three, for once more, these scaling rules do not work alone. In contemporary toxicology, *in vitro* assays even on simple properties such as solubility and permeability, as well as knowledge on the structure of the compound (number of hydrogen bond donors/acceptors, etc.), can serve to identify classes of compounds for which more specific scaling rules have been devised. Of course, a difficulty with Bernard's method is that it is unknown what is relevant in any given particular conditions: the special arrangement, in the horse, of the nostrils and the epiglottis, are noticed only after the animal's death. But this makes his science a progressive research programme. Part of the value of scaling rules in toxicology is to make departures from them stand out as targets of explanation, and thereby allow the improvement of the whole set of rules. As Bernard emphasized in the earlier quote, this epistemic iteration does not necessarily involve human observations: probability of departures from scaling rules can generally be estimated from departures in other species. Even *in silico* toxicology models, in the best of cases, are used in conjunction with animals studies: for instance, the ideal application of Hoffmann-La Roche's physiologically-based pharmacokinetic (PBPK) model involves, for each compound, a test of the model's prediction for rat in order to detect compounds which might depart from the model's rules (see for instance the pipeline in

Theil et al. 2003, p.42).

4.2.2 Theoretical constructs

Bernard's foundational philosophy made medicine scientific in the same moment that it provided a rationale for animal models, because it allowed his science to say something about organisms that differed – be they two human individuals, or a man and a frog. In doing so, it made a science that was itself constructed around no specific material system. A very similar strategy was used almost a century later in the field of psychology. Strictly speaking, genuine psychological categories are inaccessible: there is no measurement procedure for depression, anxiety, or love. These are instead inferred from accessible things, typically observable behaviours, biological correlates, or subjective reports. Because this lack of immediacy is so acute in psychology, it has prompted much more theorizing by psychologists. This discussion can be useful even for biomedical research, for the same issue is present in biology, however more subtly.

American psychologists of the first half of the 20th century adopted behaviourism[7] in response to what was perceived as speculative discussions of inaccessible mental states. However, discarding mental states altogether prevented the study of what seemed central psychological questions: "Eventually behaviourism failed, in part because it could not satisfy the need for a realistic and useful psychology of human action and thought." (Mandler 2002, p.340) The tension remained (and remains today) as to how test results can bear on anything more than the test itself – an inference sometimes dismissed as "pure speculation" (Anastasi 1950, p.67). An influential article published in 1955 by Lee J. Cronbach and Paul E. Meehl directly addressed these concerns.

Inspired by discussions in the Committee on Psychological Tests (1950-54) of the American Psychology Association, Cronbach and Meehl (1955) distinguished different kinds (or aspects) of validity of tests or measurement procedures which then became the topic of much discussion in psychology: predictive validity, concurrent validity, content validity and

[7] In a nutshell, behaviourists took psychology as a science of behaviour rather than a science of the mind, and held the doctrine that "[b]ehavior can be described and explained without making ultimate reference to mental events or to internal psychological processes" that are unobservable (Graham 2010, p.2). Burrhus Frederic Skinner (1904-1990) was probably the most famous and most radical representative of behaviourism, extending it to a social doctrine.

construct validity[8]. Importantly, all these kinds of validity are external[9]: they do not pertain to the repeatability of the test's results, but to the relationship between the result and something else – another measurement (predictive or concurrent validity) or a more abstract meaning (content and construct validity). Construct validity was certainly the Committee's most important contribution[10], and in fact later sets of criteria generally retain predictive and construct validity, to which Willner (Willner 1984, 1991), for instance, added face validity (phenomenological similarities). However, construct validity is also both the most debated concept, and the most debated aspect of a test, and for this it is also the most discussed by the authors. "*Construct validation* is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'." (Cronbach and Meehl 1955, p.282) Construct validity was an attempt to preserve at the same time scientificity and the talk about unobservables. Construct validity is what justifies (if anything) the use of tests such as the elevated plus maze (introduced in section 2.4.2) to say something about anxiety.

In order to justify construct validity, Cronbach and Meehl rely on the notion of 'nomological net', inspired from the deductive-nomological account of scientific structure and explanation (for instance Hempel 1965). For this reason, their elaborate discussion is often dismissed as based on an obsolete philosophy. As in the case of Bernard, I believe that Cronbach and Meehl's reasoning is largely independent from the nomological language, and

[8] As several different versions of these concepts have been proposed in the literature, I will not discuss the first three in detail. In a nutshell, concurrent validity measures the correlation of the test with another, previously validated test. In contrast, predictive validity assess the correlation with a future event. Content validity assesses whether a test measures all aspects of what it is meant to measure. Several authors have proposed even more refined kinds of validity, especially in the context of psychiatric diseases, for instance allowing phenotypic modeling despite differences in etiology or vice versa. Perhaps the most detailed such partitioning of validity is presented by Belzung and Lemoine (2011). Because these distinctions fall within a surrogate view of models, I will not discuss them here, but they are certainly useful to an evaluation of the surrogate role of biomedical models.

[9] Validity is not, therefore, a property of the test itself, and it is generally understood in practical terms: "validity can be assessed only in relation to the broader objectives of the research program." (Willner 1991, p.1) Samuel J. Messick, a very influential figure in the assessment of psychological tests, seems to go even further and includes what in medicine is generally referred to as 'clinical *utility*': "validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as potential – consequences of score interpretation and use (i.e., construct validity conceived comprehensively). [...] As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use." (Messick 1995, p.742)

[10] Cronbach and Meehl refer to it as "the chief innovation in the Committee's report". They note that "This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers)" (Cronbach and Meehl 1955, p.281).

is tied to what I will discuss in this chapter.

According to Cronbach and Meehl,

> "A necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least *some* of whose laws involve observables. [...] 'Learning more about' a theoretical construct is a matter of elaborating the nomological network in which it occurs, or of increasing the definiteness of the components." (Cronbach and Meehl 1955, p.290, original emphasis)

This is roughly in line with logical empiricism, which required of theoretical terms that they be anchored to statements about observables. A possible departure from (at least some versions of) logical empiricism is that talking of a net, only 'some' of whose laws involve observables, does not fit the hierarchical structure of reduction that is typical of logical empiricism. Instead, I would argue that Cronbach and Meehl are closer to a kind of coherentism *à la* Quine. Following Duhem, Quine famously argued that "our statements about the external world face the tribunal of sense-experience not individually but as a corporate body" (Quine 1951, p.41). Quine proposed a principle of 'minimal mutilation' according to which our network of knowledge is amended upon incoherences (for instance upon evidence contradictory with some of predictions of net). He suggested different values according to which amendments could be weighed. In a similar way, Cronbach and Meehl (1955) argue for a notion of parsimony, according to which additional constructs are justified if they do things (e.g. have explanatory power), although leaving some space for burgeoning constructs. "As research proceeds, the construct sends out roots in many directions, which attach it to more and more facts or other constructs." (Cronbach and Meehl 1955, p.291)

Although seldom (if ever) nomological, the relationships between different biological concepts, phenomena, or theories are fundamental in biology. The more there are such relationships, the more stable the network. What I will try to emphasize in this section is that operationalization is not only a means of studying theoretical things. The other side of the coin – already suggested by the previous section on Bernard – is that theoretical constructs are also means of bridging material systems – of studying one material system through another.

In other words, theoretical constructs can do the very thing for which the famous behavioural psychologist J. B. Watson chose to forsake them. Mandler (2002) points out

that for Watson, behaviourism served as a justification of the use of animal experimentation to study human psychology, because it allowed the "unification of human and nonhuman behaviors into a single object of investigation" (Mandler 2002, p.340). Making different behaviours refer to the same construct justifies animal modeling while preserving species differences.

4.2.3 Theoretical constructs as bridges

It is common to think of instruments and experimental systems as means of studying theoretical entities. This is most characteristic of contemporary physics, which works with unobservable entities – think for instance of the recent experiments at the CERN on Higgs' boson (see Della Negra et al. 2012 for an overview). But this subordination (experiments and instruments being means to talk of the theoretical) hides another aspect of theorizing, most important in biomedical research. Theoretical terms, entities and relationships are also means rather than ends in themselves: they are bridges between material systems.

This is, in a way, the essence of abstraction: by stripping it of particular conditions, a claim can be made to apply to a whole class of objects or instances. In this respect, what the animal modeler does is the same as what the medical scientist does: medical cases are all atypical but nevertheless grouped under specific pathological categories[11], which "although they originate from some actual observed instance in the first place, once they begin to be disseminated and used, they become idealized away from particular details of the observed phenomena." (Ankeny 2007, p.52; see also Ankeny 2010) Ankeny (2007) has argued that model organisms represent the same idealization in the context of cross-species abstraction.

Such abstractions however represent only one way, and a particularly simple one, of using the theoretical to bridge different material systems or situations. As both examples of recent toxicology (end of section 4.2.1) and post-behaviorist psychology (section 4.2.2)

[11] Ankeny (2007) discusses this logic using the notion of an 'index case', beginning with the observation of a first patient (case): "The case's use occurs through retrieval when a practitioner is presented with a new case which seems to have some overlap with the original index case, [...] The result is a feedback loop between processes of justification of the fit between the original index case and the new case under examination, particularly via assessment of similarity and identity relations. [...] new cases can lead to modification of the index case as appropriate over time, or even adoption of a new index case for a particular condition, which in turn is disseminated through publication and teaching." (Ankeny 2007, p.51)

have shown, theoretical constructs most often work in a more complex way. This was acknowledged in the NRC's Committee on Models for Biomedical Research which organized its framework around the notion of a General Biology, understood as follows:

> "the results of biomedical research can be viewed as contributions to a complex body, or matrix, of interrelated biological knowledge built from studies of many kinds of organisms, biological preparations, and biological processes at various levels." (Committee on Models for Biomedical Research 1985, p.2)

The innovative aspect of their proposal is to recognize what they call "many-to-many modelling to the matrix of biological knowledge" (Committee on Models for Biomedical Research 1985, p.3), and "encourag[e] investigators and managers to think of models not necessarily as analogs relating directly to humans on a one-to-one basis" (Committee on Models for Biomedical Research 1985, p.6). The idea I wish to retain here is that biomedical models may connect to humans not in a direct one-to-one mapping, but through a broader and more elaborate theoretical framework. The failure of nomological approaches to theories should not lead us to abandon the theoretical altogether, but to instead look for other ways in which representations and representational spaces can be made to stand in defined relations to each other. In the remainder of this chapter, I will show the implications of this in concrete debates in the biomedical sciences.

4.3 Cancer stem cells between theory and operations

In the previous chapter, I presented the example of xenograft systems in cancer research (see section 3.2.3). In this section, I will discuss such cases in more depth, especially the recent applications of xenotransplantation within the Cancer Stem Cell (CSC) framework. The aim of this discussion is to understand the relationship between theoretical terms – such as I will argue is the notion of 'cancer stem cell' – and concrete biomedical models.

By offering a more precise target, the CSC framework promised to revolutionize cancer treatment, and especially to explain and prevent relapses. However, I will argue in the following sections that the notion of CSC and its operationalization in the laboratory ought not be understood and assessed directly in relation to these clinical endpoints, but should

instead be understood in relation to the theoretical framework to which it belongs[12].

4.3.1 The theoretical grounding of early cancer xenografts

As mentioned in the previous chapter (section 3.2.3), tumor transplantation has been instrumental for cancer research since the beginning of the 20th century, and early on a number of scientists were using transplantation as a measuring procedure – as a means of grading tumours according to some value (see for instance Loeb 1945, Towbin 1951, Greene 1952). I however left unspecified the nature of this 'value'. In fact, using transplantation as a measuring procedure implied the invention of something that it measured: something whose meaning was not limited to operations – an abstract quantity – and to which transplantation provided access. Some, following Loeb, took this abstract quantity to be the 'growth momentum' of tumours:

> "Clinically, the growth momentum of a tumor, i.e., the rate of enlargement, infiltration, and metastasis, characterizes the degree of malignancy of a neoplasm. It has been shown experimentally with animal tumors that growth momentum is likewise one of the most important factors governing transplantability, particularly heterologous transplantability. [...] Accordingly, the determination of heterologous transplantability of a tumor would provide a measure of its growth momentum and, hence, the degree of malignancy." (Towbin 1951, p.716)

As I argued in section 3.2.3, this is a clear example of biomedical models used as measuring devices. The 'growth momentum' was the unobservable value to be measured (the equivalent of temperature), and the success of the transplantation was the signal (the equivalent of the height of the mercury column). For others, transplantability was instead a proxy to another property, such as the 'autonomy' of the tumours. In both cases, these abstract quantities were already loaded with both conceptual content and experience. Indeed, the notion of autonomy was already used to explain both developmental processes and carcinogenesis (see for instance the work of Hugo Ribbert or John George Adami, where both are explained as differential responses to 'tissue tension'). Similarly, the notion of 'growth momentum' was central to Leo Loeb's (1869-1959) biology (see especially Loeb 1945). Loeb himself was using transplantation as a measuring procedure, although his approach

[12] Important parts of this section were previously published in Germain (forthcoming b).

was more complex[13]. Importantly, in both cases the abstract notion was used to explain a variety of phenomena, both natural and artificial, both normal and pathological. Because it took cancer and physiology as variations of the same causes, this approach was part of what Michel Morange called "the Regulatory Vision of cancer", according to which "cancer was conceived as a disease of development" (Morange 1997, p.6). This is in contrast to the model that later prevailed, especially after the 1980's, with strictly pathological notions such as 'tumorigenicity' (until the relatively recent resurfacing of a regulatory vision, discussed in section 4.3.3).

Given the diversity of transplantation procedures (host species, site of transplantation, assessment method, etc.), it should not come as a surprise that scientists produced different, conflicting classifications. This prompts the question of which transplantation system correctly tracks growth momentum (or the abstract quantity of choice) – a question which bears strong resemblance to a core problem of classical thermometry: without a direct access to temperature, how can one know which thermometer gives the right temperature?

Before looking briefly at the analogous case of thermometry (section 4.3.2), it is important to note that both of these questions make the important assumption that there have to be single, true classification systems – or, for that matter, a single, true temperature. As Hasok Chang (2004) has argued, this is mostly explained by the fact that the notion of temperature already had an entrenched ancestry. It built upon both a long philosophical tradition that had already pre-conceptualized the notions of heat and 'caloric', and an immediate, daily-life experience of variations in temperature. Both of these loosely fitted the readings of thermometers, strongly suggesting that all were related to a common quantity. To some extent, similar arguments can be made in the case of malignancy. In the quote from Towbin above, it is striking that the term 'growth momentum' is associated to a disarraying variety of phenomena which obviously *could* point in different directions. But it was sufficient that they would align most of the time to postulate a common cause – after all, most if not all scientific laws are *ceteris paribus*.

[13] Loeb considered his experiments to simultaneously measure different aspects of the phenomena which he called 'differentials' (Loeb 1945). Although I believe that the present discussion could equally apply to his work, the presence of multiple quantities in the same reading makes the matter less straightforward. Beside his achievements in tissue transplantation, Loeb is known for his work on cancer and on endocrinology. This juxtaposition is not trivial, and allowed him to study both malignant and normal growth in a unified biological picture.

Moreover, the clinical context of xenotransplantation experiments already provided a very concrete notion of malignancy: the clinical outcomes of patients provided a grading of tumours. There are different ways of looking at this clinical reality, and a first one is to consider the clinical as the (*de facto*) unobservable to which the instruments are giving access: biomedical sciences should after all allow the prediction of clinical outcomes, such as the effect of intervention. In this context, transplantation experiments simply ought to approximate clinical outcomes. Indeed, interesting correlations were observed early on, for instance the fact that "all tumors that manifest the ability to grow as isolated satellites of the primary lesion are also capable of growth in alien hosts" (Greene 1952, p.43)[14], and scientists tinkered their transplantation procedures to approximate clinical knowledge. Nevertheless, scientists did not go all the way in this direction. Beside issues of sensitivity[15], two important reasons can be given for this. The first is that these scientists were not just after a predictive system, but an explanatory, or at least exploratory one, enabling an investigation of the "mechanisms of autonomy"(Greene 1951, p.902)[16]. In this context, approximating the clinical outcome has the value of highlighting departures from it, and therefore of stabilizing these departures as objects of explanation.

A second reason, more important for our purposes, is the lack of repeatability of what we might call the 'clinical measurements' of malignancy. Actual patients are all atypical in some respect, especially given the fact that the disease's index case is idealized (Ankeny 2010). Malignancy, understood as the ability for pathological, neoplastic growth, is a relational property of cells (if not of tissues – see Sonnenschein and Soto 2008) with respect to an environment. Indeed, cells can to some extent become malignant just because of differences

[14] And because metastasis is so problematic from a therapeutic point of view, transplantability also meant a bad clinical outcome: "It will be noted that sixty-one, or 93.8 per cent, of the patients whose tumors proved to be transplantable are now dead and that four, or 6.1 per cent, are living. In contrast, forty-six, or 79.3 per cent, of the patients whose tumors were not transplantable are still alive; and only twelve, or 20.7 per cent, dead." (Greene 1952, p.31)

[15] The test was used in some laboratories (Towbin 1951), but its sensitivity was criticized. Hence in the decade that followed, different methods were shown to make the hosts more receptive, including X-ray irradiation, cortisone treatment, and thymectomy, but the most important advance was certainly the discovery, in the early 1960's, of the *Nude* mouse mutant. In fact, even normal tissues successfully engrafted on the *Nude* mouse, but at that time Greene's idea of transplantability as a test of cancer was already forgotten.

[16] See also the quote from Greene discussed in the previous chapter: "The fact that a biological quality as fundamental as the ability to grow in an alien species differentiates morphologically identical tumors suggests that the tumors must also differ in metabolic or biochemical constitution. It would seem important, therefore, to distinguish tumors with respect to this property and to study the different groups formed rather than to consider morphological similarity a proof of constitutional identity." (Greene 1952, p.41)

in the surrounding tissue (Bhowmick and Neilson 2004), and alternatively can be made non-malignant in the appropriate environment (see for instance Allegrucci et al. 2011). This means that although each human tumour is associated to a medical history and clinical outcome, this history has many determinants that are external to the cancer cells. These can be due to the exact site and micro-environment of the tumour, the patients' constitution or genetic background, treatment history, etc., so that the correlation between the nature of the cancer cells and the medical outcome is messily statistical rather than deterministic. Nevertheless, experience strongly suggested that among the factors of malignancy were differences that were intrinsic to the cancer cells: hence the invention of abstract notions such as 'growth momentum' served to track these factors. In the absence of an independent means to group tumours together (which transplantability was trying to offer), this meant that in practice each clinical outcome was yet another poor approximation of the tumour's (or cells') intrinsic malignancy (i.e. growth momentum, autonomy, etc.). In this view, the clinical outcomes are analogous to the pre-conceptualized notion of temperature: they are a first accessible but limited readout of something more robust. The invention of such an intrinsic property, because it makes it transportable (transplantable), enables the phenomena to simultaneously become a theoretical variable and an object of experimental study.

4.3.2 Epistemic iteration

To study cancer in a mouse, moreover in the highly artificial context of the laboratory, will never be the same as to study it in patients. However, the invention of an abstract property to which the measurement procedure would give imperfect access enabled the use of the mouse system to study the human system. In other words, the abstract concept (e.g. growth momentum, autonomy) provided a bridge between different material systems, by assuming that the material systems were simply two imperfect operationalizations of the same theoretical construction – or, as Bernard would have put it, the same theoretical principle in different starting conditions. To understand the use of this detour, it is helpful to look again at the analogous example of thermometry.

A central problem of thermometry was to reconcile the fact that many different ther-

mometers (air, mercury, etc) gave very different (not linearly correlated) readings. In the end, most of the conundrum was solved when Thomson (Lord Kelvin) reasoned that the establishment of an absolute temperature required "a theoretical relation expressing temperature in terms of other general concepts" (Chang 2004, p.175), which he took from Carnot's practical model of heat engines:

> "As Thomson was attempting to reduce temperature to a better established theoretical concept, the notion of mechanical effect (or, work) fitted the bill here. A theoretical relation between heat and mechanical effect is precisely what was provided by a theory of heat engines." (Chang 2004, p.175)

The first step was therefore to postulate an abstract temperature as defined by its theoretical relationship with another abstract term, 'work', which was linked to operational concepts through mechanics. Interestingly, the second step was what Chang describes as a "deliberate conflation" of this absolute temperature and of the temperature given by any thermometer: physicists assumed that the thermometers gave imperfect readings of this abstract quantity, and simply substituted one for the other in their formula (Chang 2004, p.214). Obviously, the fit was not perfect, but discrepancies allowed scientists to recalibrate their instruments, and engage in successive steps of approximation and recalibration which Chang characterized as 'epistemic iteration': "point-by-point justification of each and every step is neither possible nor necessary, what matters is that each stage leads on to the next one with some improvement." (Chang 2004, p.215). Concretely speaking, the most important output of this iteration was the coefficient of thermal expansion of different substances. In other words, it was a means of understanding (explaining as well as predicting) departures of any thermometer from the theoretical (or absolute) temperature.

It is interesting to note that the theory provided both the motivation and the solution to the problem: scientists investigated thermometry to build a theory of temperature, and yet it is the theory that ultimately solved the problem of thermometry. Or perhaps dissolved the problem: by providing ways of going from one thermometer to the other, it undermined the need for 'the right thermometer'.

The value of theoretical notions is thus that they come with relations and implications. The conscious conflation of such a theoretical notion with one of its operationalization therefore has implications for the operational context. These implications will, or will not

obtain. But the point is that departures from the theoretical notion call for an explanation, and these explanations feed back as corrections or *ceteris paribus* clauses into the relationship between the two. These are simultaneously informative about both the theoretical and the operational notions. The fact that these corrections are relative to only one operationalization at a time makes them experimentally tractable. The coupling between operations and theory is productive if it leads to a tightening of their relations.

4.3.3 The Cancer Stem Cell framework

At the turn of 2000's, strong analogy between physiology and pathology resurfaced in cancer research under the form of the Cancer Stem Cell (CSC) explanatory framework. In a widely cited paper published in 1997, Dominique Bonnet and John E. Dick showed that (in acute myeloid leukemia) only a small proportion of leukemic cells are capable of initiating cancer in immunocompromised mice, and that these cells show many characteristics of normal hematopoietic stem cells: most importantly the potential for self-renewal and differentiation, as well as similar cell-surface markers[17] (Bonnet and Dick 1997). This implied that tumours, like healthy tissues, are hierarchically organized – a point which was developed by further studies (see for instance Hope et al. 2004). It also suggested that only a small proportion of the cancer cells were responsible for the sustained growth of the tumours. This had major clinical implications: it meant that measurements of clinical progress such as tumour growth or shrinkage were potentially misleading, and that instead of fighting the bulk of the tumour one should concentrate on a specific subpopulation. It also offered an explanation to the daunting problems of cancer recurrence and resistance to therapy, thereby fuelling hopes of a solution. This quickly propelled the CSC hypothesis to the forefront of cancer research.[18]

The discovery of leukaemic stem cells prompted a whole research programme aimed at identifying and isolating CSC in other forms of cancer, starting especially with breast

[17] Fluorescence-activated cell sorting (FACS), and more generally antibodies, were indeed a connecting point between pathology and physiology. On this topic, see especially Cambrosio and Keating (1995) and Cambrosio and Keating (2003).

[18] Some of the continuities and departures of the CSC framework from the previous 'oncogene paradigm' are discussed in Blasimme et al. (2013). The relationships between the CSC model and clonal evolution by natural selection within tumours are explored in Germain (2012). For an interesting scientific review of the CSC model, see Visvader and Lindeman (2012), Clarke et al. (2006), or Valent et al. (2012).

cancer (Dick 2003) and brain tumours (Singh et al. 2003). There was however a wide heterogeneity in both experimental assays and conceptual tools. Most fundamentally, the question of whether cancers were hierarchically organized was often conflated with the question of the 'cell-of-origin': whether cancer initiation took place in stem cells, i.e. the CSC were originally transformed stem cells, or instead de-differentiated cancer cells. In an attempt to clarify these issues, to standardize both language and assays, and to evaluate the published evidence, the American Association for Cancer Research sponsored a workshop gathering the pioneers in the field (Clarke et al. 2006)[19]. The workshop entrenched the expression 'cancer stem cell', rather than alternatives judged more ambiguous, such as 'cancer-initiating cell', and offered a definition of the concept:

> "The consensus definition of a cancer stem cell that was arrived at in this Workshop is a cell within a tumor that possesses the capacity to self-renew and to cause the heterogeneous lineages of cancer cells that comprise the tumor. Cancer stem cells can thus only be defined experimentally by their ability to recapitulate the generation of a continuously growing tumor." (Clarke et al. 2006, p.9340)

The definition explicitly avoids any conflation of the cancer stem cell hypothesis with the cell-of-origin hypothesis (a demarcation reiterated in later meetings – see for instance Valent et al. 2012, p.769). It however contains its own ambiguities, which have since been central to the discussion on CSC. Note the language of the first sentence: a cell that 'possesses a capacity' and 'causes' some things. The language of capacities is well-chosen, for what is at stake here is precisely Cartwright's metaphysics of causation (Cartwright 1989)[20]. The question is whether given cells have a tendency toward given behaviours not in a specific

[19] The workshop report summarizes the CSC hypothesis in the following way: "In the cancer stem cell model of tumors, there is a small subset of cancer cells, the cancer stem cells, which constitute a reservoir of self-sustaining cells with the exclusive ability to self-renew and maintain the tumor. These cancer stem cells have the capacity to both divide and expand the cancer stem cell pool and to differentiate into the heterogeneous nontumorigenic cancer cell types that in most cases appear to constitute the bulk of the cancer cells within the tumor. If cancer stem cells are relatively refractory to therapies that have been developed to eradicate the rapidly dividing cells within the tumor that constitute the majority of the nonstem cell component of tumors, then they are unlikely to be curative and relapses would be expected. If correct, the cancer stem cell hypothesis would require that we rethink the way we diagnose and treat tumors, as our objective would have to turn from eliminating the bulk of rapidly dividing but terminally differentiated components of the tumor and be refocused on the minority stem cell population that fuels tumor growth." (Clarke et al. 2006, p.9339)

[20] Cartwright (1989) argued that "[t]he generic causal claims of science are not reports of regularities but rather ascriptions of capacities, capacities to make things happen, case by case." (Cartwright 1989, p.2-3) Capacity claims represent what it is in the nature of an object or property to cause (see Cartwright 1999, p.72). The application of Cartwright's metaphysics to biological phenomena is tricky insofar as entities and properties are most often neither stable, nor precisely bounded. Cell types or states, such as being a CSC, represent an excellent example: it is often unclear whether a given cell is of a given type or not, and

context, but robustly across a set of contexts – and this is a precondition of the whole experimental approach.

Most often, cancer stem cells are operationally defined by their ability to grow a tumour (recapitulating the heterogeneity of the original tumour) when serially transplanted in an immunocompromised mouse. Obviously, what matters clinically is the ability of the cells to sustain tumour growth not in a mouse, but in the human patient. The latter is, however, experimentally inaccessible: while there is correlative evidence, one cannot establish causality in human patients because one cannot conduct controlled interventions testing the ability of cells to grow in humans. Even if it were ethically permissible, there would be more practical epistemic hurdles to such experiments: to give a simple example, some cells will initiate tumours in some micro-environments of the body but not in others. Generally, the clinically relevant question is whether given cells of a specific tumour, in a post-treatment micro-environment, would sustain the growth of a new tumour (and hence relapse). If yes, they must be destroyed. However, no experimental system other than the specific clinical case will accurately predict this. More importantly, even if there was such a system, its usefulness would be limited to the specific patient. What one can expect of an experimental system, therefore, is to make robustly visible a defined dimension of the clinically-relevant phenomena.

As in the previous case, an important strategy in this context is to shift attention to a theoretical entity or concept. Because it has theoretical ties to a series of other concepts which are more easily operationalized, it is empirically grounded. At the same time, because the concept is not a highly contextual operational concept, it can bear on an experimentally inaccessible setting (i.e. the clinical case). As the report of the 2011 Working Conference on CSCs explicitly discusses, the notion of CSC is a theoretically defined one, strongly related but distinct from any single operationalization (Valent et al. 2012, see especially table 1 on p.770). This is generally perceived as a problem to overcome: one should replace the theoretical and blurry concept with an operational one. The point that I wish to make is that doing so would rob the concept of its utility: if CSC-ness is defined as, say, a particular procedure of serial transplantability within a specific host, then CSC-ness cannot

even when it is clear, they can change state according to different triggers. Nevertheless, much of biology relies on ascription of capacities to cell types.

at the same time be defined as sustaining tumour growth in the patient, and hence the CSC framework cannot fulfill its promises. Instead, one should precise the meaning of the term without restricting it to an operationalization. A theoretical term is not vacuous or imprecise merely because it is theoretical, but can be made precise insofar as it entertains tight connections (e.g. implications) to other terms.

4.3.4 Melanoma-initiating cells

To see the implications of this point, it will be useful to present the specific example of a debate as to whether melanoma followed a CSC model. The experimental paradigm of the CSC framework is to divide the tumour in subpopulations of cells according to some markers (typically on the cell's membrane, so that cells can be sorted through antibody-based methods such as FACS), and assess whether these subpopulations differed in terms of some measurement. In the field of melanoma research, scientists have proposed to speak of melanoma-initiating cells (MIC) as an operational definition of CSC: MIC are cells which, when serially transplanted into an immunodeficient mouse, are able to produce tumours recapitulating the heterogeneity of the original tumour. Strictly speaking, scientists are most often not measuring whether the cells are able to produce tumours, but to what extent, and therefore the injected cells are not said to be all CSC, but to be enriched in CSC. Once more, mice were recruited as measuring devices, and once more, a variety of transplantation procedures resulted in conflicting measurements.

A few years ago, Schatton et al. (2008) identified a sub-population of cells, ABCB5+ cells (cells expressing the ABCB5 antigen at their surface) enriched in what they claimed to be CSC. In order to test it against the operational definition of MIC, they transplanted ABCB5- and ABCB5+ populations of cells from a human tumour into NOD/SCID immunodeficient mice ('Non-Obese Diabetic/Severe Combined ImmunoDeficiency') and looked at the tumour progression. After 8 weeks, hardly any tumour grew in the first case, and the majority steadily grew tumours in the second group. In other words, only a small proportion of tumour cells, strongly enriched in the ABCB5+ population, were able to initiate and sustain new tumours. They published an enthusiastic letter to nature which was heavily cited, and for a time it was proclaimed that CSC had been identified in melanoma.

Some months later, the lab of Sean J. Morrison's lab (Quintana et al. 2008) published a paper attacking these claims. The most important for the present discussion is that they tried the same experiments with an even more immunocompromised mouse (the NOD/SCID Il2rd-/- mouse) and obtained radically different results. Injecting single cells, they found that one out of four was able to initiate palpable tumours, and trying a wide range of markers, they were not able to correlate this with any signature. They therefore concluded that evidence argued against the adequacy of the CSC model for melanoma, and that experiments seeking MICs should beware of relevant differences between the tumour environment in the patient and in the mouse host. The lesson, it seems, was that Schatton et al. (2008) had drawn a bad conclusion that was due to the particular mouse model they used, which happened to be unrepresentative of the human host. Quintana's paper was (and still is) a big success, being cited even more than the first, often as a methodological warning. Nevertheless, given how 'unnatural' the dramatically immunocompromised mice are, there is still considerable debate as to which is the best surrogate for the patient (see for instance Civenni et al. 2011).

The mice, and in fact the whole experimental system, again acted as an instrument: they transformed an unobservable, yet causally relevant difference, into a visible signal, thus revealing this difference. But what difference exactly? Given the disagreements of two recipients, which signal faithfully informs us about cancer-initiating or 'CSC-ness'?

The question becomes even more acute if we consider the rest of the story. Slightly more than a year after Quintana's paper, Schatton et al. (2010) published a follow-up paper in *Cancer Research*, apparently moving the topic: "Modulation of T-Cell Activation by Malignant Melanoma Initiating Cells". Taking the discrepancies between the two studies as a starting point, they addressed the question of why the difference between the mouse strains – the absence or presence of the interleukin-2 gamma receptor (Il2rg) – made such a difference to the apparent role of ABCB5+ cells. It turned out that ABCB5+ cells seem to block or reduce the proliferation of immune cells and the production of interleukin-2, thus modulating T-cell activity. Obviously, in a mouse that anyway lacks such an activity (and, as a matter of fact, completely lacks interleukin-2 gamma receptors), one expects to find no difference between the subpopulations of cells. But in a mouse that has such an

activity, only cells that are able to disrupt this mechanism can proliferate efficiently.

Assuming that ABCB5+ cells prove to also be more malignant in the case in humans, one might argue that the first model (the least immunocompromised) was a better model. However, Morrison's group would probably point out that this malignancy is not due to the tumorigenicity of the cells *per se*. But *on what ground can one exclude phenomena as part or not of such an abstract property*? On closer inspection, which instrument is the best depends on what it is that we wish to measure – in this context, on the understanding one has of 'cancer-initiating' or CSC-ness.

MICs have the capacity to initiate the sustained growth of tumours, but in the presence or in the absence of an immune pressure? One the one hand, human tumours do not develop 'in the void': cancer patients are seldom so immunodeficient, and immune response is an important part of cancer development and of variability in outcomes. A notion of tumorigenicity independent of this pressure seems to be an idealization that lacks practical relevance. On the other hand, it seems scientifically worthwhile to isolate the different components influencing the malignancy of cancer cells, so that we might want to exclude the effects of the immune system: tumorigenicity is one thing, evasion of immune surveillance is another. The problem with this reasoning is that many other causally relevant elements (many ways through which some cells might be more tumorigenic, or initiate cancer more than others) could also be excluded. Therefore, one can legitimately ask why excluding this and not other causally relevant elements. The only reasoned answer one can provide has to be linked to the adoption of a theoretical framework[21] – or what one could call the 'theoretical grounding' of operations or instruments.

In the first paper that I briefly described in the previous section, Quintana et al. (2008) reduce CSC-ness to tumorigenicity: "the cancer stem-cell model has suggested that only small subpopulations of cancer cells have tumorigenic potential" (Quintana et al. 2008, p.593). There has been a general tendency, especially in the field of melanoma, to avoid the abstract talk of CSC in favor of the operational talk of melanoma-initiating cells – or cells that initiate melanoma when transplanted into an immunodeficient mouse. Likewise, participants of the 2011 Working Conference on CSC seemed very concerned by conflations of the conceptual and operational definitions (Valent et al. 2012). However, severing the

[21] On this point, see also Griesemer (1992).

connection between the two is equally problematic. An exclusive focus on tumour-initiating potential would be like a focus on the height of the mercury column: while it might be useful locally, it does not allow semantic extension. The abstract concept does. The CSC framework can potentially mediate between the material contexts by offering a constrained rationale for the different contextualizations of the same concept. But this means that the problem of selecting the 'right' xenograft model can only be solved if one has at least a tentative theoretical understanding of what it is that the instrument should measure.

The CSC framework can succeed where the notion of growth momentum has failed precisely because its meaning has theoretical implications which are not reducible to the operational definition of CSC. The seminal findings of Bonnet and Dick (1997) was not that some leukemic cells were more tumorigenic than others, but precisely that those were the cells possessing stem-cell like characteristics ("the differentiative and proliferative capacities and the potential for self-renewal" Bonnet and Dick 1997, p.730). In doing so, it established a parallel between cancer and normal development, suggesting that the physiological differentiation hierarchy can shed light on the dynamics of cancer. In other words, it also poses additional constraints as to the kind of measurements that ought to be linked to it. Schatton et al.'s findings of the modulation of T-cell activity by cancer cells may be extremely relevant for an understanding of cancer, but it has no physiological counterpart and is unrelated to the tissue hierarchy. As such, it is irrelevant to the identification of CSC. It is not biologically or clinically irrelevant, but irrelevant to what it is that the xenograft was supposed to measure. This point is often missed by arguments focusing on similarity between the model and the patient: for instance, following on the melanoma debate, Civenni et al. (2011) repeatedly insist on the xenograft having to be "exact phenocopies of the parental tumors" (Civenni et al. 2011, p.3100; see also DeRose et al. 2011). An obsession for mimicry is dangerous not only because ultimately it will always fail, but also because it neglects the function of the model. Insofar as the xenografts are used for the identification of CSC, the question of the 'right' assay does not necessarily hinge on the similarity of the (grafted and parental) tumours, but can only be answered with respect to the theoretical meaning of CSC: how does the measurement relates to the theoretical implications of the CSC framework? What are the structural relationships between the space of representation

offered by the measuring device, and those in which we represent development?

4.4 Spaces of representation

Biomedical models should not be seen as aiming to replace humans, but as means of establishing a space of representation that can be related to other spaces through theoretical relationships. Stepping back from the concrete examples discussed in this chapter, I would like to situate these claims with respect to a more general coherentist account of science, and to the notion of epistemic iteration discussed earlier.

I will begin with Rheinberger's position on representation:

> "I argue that there is no such thing as a simple representation of a scientific object in the sense of an adequation or approximation to something 'out there,' neither conceptually nor materially." (Rheinberger 1995a, p.114)

Rheinberger subscribes to Derrida's point about the erasure of the signified in front of its signifier: against platonic ideas of an absolute 'signified' being gradually corrupted by its representation, Derrida has argued that the signified is always itself a signifier – that the signifier refers not to some transcendental signified, but to more signifiers (see for instance (Derrida 1967)). Rheinberger applies this to experimental sciences in the following way:

> "I have posited that with an experimental system a space of representation is established for things that otherwise cannot be grasped as objects of epistemic action. Biochemical representations, to use an example, create an extracellular space for reactions assumed to take place within the cells. Conventional wisdom has it that such a representation constitutes a model of what is going 'out there in nature.' Thus, biochemical *in vitro* systems are taken to be models of *in vivo* processes. But what goes on within the cells? There is no other way to know this than to provide a model for it. In other words, nature itself only becomes real in a scientific and technical sense as a model. Of course, there are *in vivo* experiments. But insofar as they are parts of a research arrangement, they are model systems, too. [...] Consequently, we have to conclude that the reference point of any experimentally controlled system can be nothing else but another experimentally controlled system." (Rheinberger 1995a, p.115-116)

Contrarily to van Fraassen (2008), it would seem that for Rheinberger spaces of representation are all material spaces: it is after all the technical conditions that, according to Rheinberger, determine a space of representation (Rheinberger 1995a, p.111). On the other hand, if this space is determined by the technical conditions, they are necessarily distinct

from them, yet their nature is hardly discussed. To see the importance of this, it is worth looking at a related problem raised by Richard Burian when commenting on Rheinberger's contribution: if each experimental systems constitute a different space of representation for something which has no absolute referent, what is the nature of their relation? Burian writes:

> "What must be explored is how one might justify the claim that scientists working with different experimental systems, and thus different *epistemic* things, might nonetheless have good grounds to hold that they had gotten hold of different parts of aspects of the same elephant." (Burian 1995, p.130, original emphasis)

It is precisely because of this issue that operationalism failed as a theory of meaning and as a foundation for science[22]: it did not allow semantic extension, or in other words it did not allow a term to designate anything else than the way it was measured, denying the possibility of a continuity across experimental settings. It is the same criticism that Burian addresses to Rheinberger.

I believe this issue is tightly related to Weber's (2005, 2011) point that an account focusing exclusively on experimental systems would miss an important dimension of science: even experimental practice depends on concepts and 'ways of acting'. Weber writes that "[w]hat New Experimentalists have tended to overlook is that experimental systems always come with theoretical interpretations of what happens in an experiment" (Weber 2011, p.218). In a similar way, Hans Radder notes that "[i]n practice, scientists designate one part or aspect of the overall theoretical interpretation of the experimental process as its result" (Radder 2003, p.157)[23]. It is the theoretical, or representational nature of the result which makes it somehow autonomous from, or reaching beyond the precise experimental settings.

4.4.1 The closure of representation

While downplaying the role of theory has been instrumental in drawing attention to major aspects of science, I believe concepts and theories must be reintroduced into the picture for

[22] Neither of which it was initially supposed to provide – see Chang (2009)

[23] Bogen and Woodward (1988) have made a similar point in the discussion of the data/phenomena distinction.

a full understanding of experimental biology. van Fraassen (2008) more than acknowledges the role of theory, indeed he has kept it at the core of his philosophy. Hence his spaces of representation are, even with concrete representations, inherently theoretical: even the operation of measurement "locates the target in a theoretically constructed logical space" (van Fraassen 2008, p.2).

A central tenet of van Fraassen's empiricist structuralism is that "[s]cience represents the empirical phenomena as embeddable in certain abstract structures (theoretical models)" (van Fraassen 2008, p.238). His account defines in precise terms (mathematical functions such as embedding or isomorphism) the relation between theory and the phenomena, however the 'phenomena' involved in this relationship cannot be the raw phenomena as chunks of reality: these are not mathematical objects, and therefore would not have a defined range for such a function. Instead, the relation is between the phenomena already represented as a structure – what is often called a data model (Suppe 1974, van Fraassen 2008).

Notice the similarity with Rheinberger's (and Derrida's[24]) point that "anything 'represented,' any referent, upon closer inspection and as soon as we try to get hold of it, is turned itself into a representation." (Rheinberger 1995b, p.50) To the philosopher of science, however, this is an unsatisfying answer which only defers the problem of representation to the relation between the data model and the phenomena:

> "the claim that the theory is adequate to the phenomena *is not the same* as the claim that it is adequate to the phenomena as represented by someone (nor as represented by everyone, or anyone)." (van Fraassen 2008, p.259, original emphasis)

And here Van Fraassen's provocative strategy is to invoke the indexicality of representation and argue that for the person who has represented the phenomena as such a model, they (the phenomenon and the data model) "are indeed the same" (van Fraassen 2008, p.259):

> "in a context in which a given model is someone's representation of a phenomenon, there is for that person no difference between the question whether a theory fits that representation and the question whether that theory fits the phenomenon." (van Fraassen 2008, p.260)

This is what van Fraassen calls a "pragmatic tautology", although one might question its tautological status: at first sight, that I take a data model to be a representation

[24] "Le signifié y fonctionne toujours déjà comme un signifiant." (Derrida 1967, p.16)

of a phenomenon does not forbid that I might consider it an imperfect representation. However, as we know from the Kantian tradition, representation is a condition of possibility of the phenomenon being an epistemic object in the first place, which means that if I doubt a data model, it must be because it conflicts with other representations I have of the phenomena. These representations are not necessarily scientific: thermometers, for instance, were evaluated early on against our bodily sensations of temperature (Chang 2004), and medical representations (such as disease categories) are evaluated against clinical outcomes which are themselves already represented *as*. This prompts a question similar to Burian's: how are we to know that two representations refer to the same phenomenon? This is, however, a realist take on the question which van Fraassen would most certainly resist, and there is another way to cast it.

In general[25], the phenomenon that is considered an object of study is not entirely new. The study of tuberculosis, for instance, stems from the study of something that was known as 'consumption'. Likewise, long before Lord Kelvin defined absolute temperature, there was a pre-scientific knowledge of temperature. How did we know that these two temperatures were the same? Because they both stood in the same set of more or less theoretical relationships to other things, from bread-baking to the making of heat engines (section 4.3.2). Obviously, the space of representation offered by thermometers interlocked with other accepted spaces of representation in ways that were not possible with bodily sensations, but most relationships in which (intersubjective) bodily sensations of temperature entered were conserved with thermometers (Chang 2004).

Even anti-realists (or agnostics) must claim some sort of adequacy between our representations and the world: a representation or space of representation is better than another if it is more *useful*. My interpretation of van Fraassen's pragmatic tautology is that this usefulness is itself assessed on the basis of representations of the phenomena as something which the theory can handle. To give a simple example, symptoms (including patients' self-report) are used both for the diagnosis and study of a disease, and for assessment of its cure: it is, therefore, the same representations that are used to build the theory and to assess its success. And perhaps paradoxically, the best illustration of this is in Hacking's

[25] Instrumental anomalies, or 'effects' (Hacking 1983, p.244), which are closely tied to a representational space, represent a notable exception.

famous statement about electrons: "So far as I'm concerned, if you can spray them then they are real." (Hacking 1983, p.23) How do we know when we have sprayed an electron? The point was made by Hacking himself in his discussion of the self-vindication of the laboratory sciences: "They are self-vindicating in the sense that any test of theory is against apparatus that has evolved in conjunction with it – and in conjunction with modes of data analysis." (Hacking 1992, p.30) The fact that instruments and theoretical frameworks mutually vindicating each other is obvious if we consider that measuring devices locate their object on a theoretically constructed space. An important implication of this is that neither instrument nor theory is given, so that in principle *both can be tinkered with*.

4.5 Conclusion

In the previous chapter, I have suggested that an important and under-appreciated usage of biomedical models is as instruments, or measuring devices. An important reason for focusing on this role was that it suggests a different view of biomedical models in which models are not aimed at mimicking or replacing the target system, but are instead *projections* of the target on a different space of representation. A space has a structure beyond what populates it, such as defined dimensions, and as such it can enter in structural relations with other spaces of representation. Measuring devices locate their object on a theoretically constructed space which is tightly constrained and well defined, while other biomedical models locate their target on spaces that are often much less defined, and therefore also much less constrained.

The relevance of such a space of representation comes from its superposition with others. This is particularly clear in the example of early cancer xenografts (sections 4.3.1-4.3.2). To start with, the space was populated with fixed points: the same tumour cells from the same tumour, taken at the same point in pathogenesis, were expected to have the same transplantation fate. However, as I argued in Chapter 1 (section 1.2.2), this sameness is neither given nor absolute, and the cells are held to be the same because they share the same position within yet another space of representation: the same patient, the same anatomical region, the same transplantation conditions, etc.

For this reason, the relationship between different models, including between patients

and animal models, ought to be understood as a relationship between the representational spaces they offer. This is what I have tried to illustrated with the example of cancer stem cells, especially the debate on melanoma-initiating cells (section 4.3.4). Instead of asking to what extent a model is 'like' another, one should ask what space of representation it offers, how this space is structured, and how it relates to others (such as our representations of development) from an inferential point of view.

Finally, once a theoretical apparatus is recruited, it must be acknowledged and harnessed to its full potential. Understanding that both the instrument and the theoretical apparatus it is related to can be tinkered with opens new possibilities in their arrangements and mutual accommodation.

Chapter 5

In vitro models and distributed modeling

5.1 Introduction

This chapter is centered around *in vitro* models, which in the last decades have been a major reason for a re-evaluation of animal experimentation (section 0.3.1). Scientists have often argued that *in vitro* models are complementary to, rather than substitutes for animal experimentation[1], and one of my aims here will be to explore the meaning of such claims. Doing so will also lead me to challenge the third assumption identified at the end of Chapter 2, namely that modeling can be understood as a dyadic relationship between the model and its target (section 2.4.3).

My main example to explore the meaning of the *in vitro* will be stem cell models based on cellular reprogramming, because they offer new possibilities which have promised to transform biomedical research (see especially sections 5.3.1 and 5.5). The first part of the chapter begins with a recent history of the cellular reprogramming technology (section 5.2.1), and discusses some of the epistemological issues surrounding stem cell models and *in vitro* models more generally (sections 5.2.2 to 5.3.2). An important point made is that, both historically and epistemologically, the *in vitro* and the *in vivo* are defined in relation to each other (see especially section 5.2.4), and the nature of this relation is explored.

[1] See for instance the contribution of Kurt Benirschke to the 1977 report of the Institute of Laboratory Animal Resources: "In my opinion, these two systems are most often complementary rather than substitutes for each other." (Institute of Laboratory Animal Resources (ILAR) 1977, p.67)

The second part (section 5.3) of the chapter discusses specifically the experimental research paradigm attached to induced pluripotent stem cells (iPSC). This will show with a rather simple case the inadequacy of the dyadic view of modeling (see especially sections 5.3.3-5.3.4). Drawing on some implications of this example, section 5.4 proposes the notion of distributed modeling as a better way to represent and understand modeling, both as a process and as a relationship. I explore some of the implications of such a view for the evaluation of biomedical models (sections 5.4.3-5.4.4), and end with some concluding remarks on the often foretold possibility of a replacement of animal experimentation with *in vitro* models (section 5.5).

5.2 Stem cell models

5.2.1 A brief history of iPSCs

In 2006, Shinya Yamanaka's lab reported the reprogramming of differentiated cells to a state of pluripotency (akin to embryonic stem cells) by inducing expression of four 'pluripotency factors' (Takahashi and Yamanaka 2006). Six years later (a surprisingly short time) he was awarded the Nobel prize in physiology and medicine, along with John B. Gurdon (mostly for his work in the 1960's on reprogramming by nuclear transfer). In their 2012-10-08 press release announcing the award, the Nobel Assembly highlighted two major ways in which cellular reprogramming was "ground-breaking". First, it "completely changed our view of the development and cellular specialisation" (The Nobel Committee for Physiology or Medicine 2012, p.1) by demonstrating that differentiation was entirely reversible – and quite easily so (only four factors!). In this respect, it is fair to say that iPSC was rather like the 'cherry on top' of many scientific discoveries since the pioneering work of Gurdon[2]. Like many others, Gurdon had long foreseen the "eventual identification of reprogramming molecules" (Gurdon and Byrne 2003, p.848). In an interview, stem cell biologist Alan Colman said

> "What we all believed... was that it could be done but must be a very, very complicated process. [...] Then Shinya comes along and showed only four factors could do it." (quoted in Vogel and Normile 2012)

[2] For an interesting discussion of the history of reprogramming and transdifferentiation, see Graf (2011).

Yamanaka and the others scientists involved did more than add a contribution to the already massive rethinking of development: they turned these discoveries into a technology. And this is the second way in which the Nobel committee praised cellular reprogramming, more specifically iPSC as constituting "invaluable tools for understanding disease mechanisms and so provide new opportunities to develop medical therapies." (The Nobel Committee for Physiology or Medicine 2012, p.3)

Again, there are two related components here. The therapeutic potential of iPSC is mostly a facilitation of stem cell therapies, first an ethical facilitation (bypassing embryos as a source of pluripotent cells), but also a potential technico-medical facilitation: stem cell therapy could now be conducted with the patients' own cells, greatly reducing the risks of immune reactions (on the ethical issues related to iPSC, see (Testa 2009, Blasimme 2012, Blasimme et al. 2013, Testa ming)). The second component is more relevant to the present work: beside promises of treatment, iPSC are also tools for studying diseases.

In their initial paper about reprogramming of mouse fibroblasts, Takahashi and Yamanaka (2006) already foresee the possible applications for stem cell therapy, and in fact make it one of the main motivations for their work[3]. Interestingly, however, they make no mention of using iPSC for the study (rather than treatment) of diseases. Research usages are mentioned almost in passing by some commentators, for instance a Nature News article states that iPSC could provide "a source of cells for research into the pathogenesis of complex diseases" (Rossant 2007, p.2), without elaborating further. This research avenue starts to become explicit in Yamanaka's second famous paper, which translates the reprogramming technology to human cells. Already in the first sentence of the abstract, they write of creating "patient- and disease-specific stem cells" (Takahashi et al. 2007, p.861). The patient-specificity is first thought within a therapeutic endeavour, (although, as we will see in a moment, it has important implications for the study of diseases), but disease-specific cells, in the sense of cells representing a disease, make sense particularly in the context of research.

[3] They begin their paper with the medical and ethical difficulties of stem cell therapy, stating that "One way to circumvent these issues is the generation of pluripotent cells directly from the patients' own cells." (Takahashi and Yamanaka 2006, p.663; see also Takahashi et al. 2007) In the same vein, the first paper demonstrating the therapeutic potential of iPSC states that "The ethical debate over 'therapeutic cloning,' as well as the technical difficulty and inefficiency of the process, has spurred the quest to achieve reprogramming of somatic cells by defined factors" (Hanna et al. 2007, p.1923).

The differentiation of iPSC into defined lineages also becomes much more important in the second paper. In the first, differentiation was limited to the very technical role of validating (or proving) pluripotency. While differentiation also plays this role in the second paper, it is not limited to this purpose but rather becomes central to the research usage of iPSC. This is emphasized in the conclusion of their paper:

> "Even with the presence of retroviral integration, human iPS cells are useful for understanding disease mechanisms, drug screening, and toxicology. For example, hepatocytes derived from iPS cells with various genetic and disease backgrounds can be utilized in predicting liver toxicity of drug candidates." (Takahashi et al. 2007, p.869)

Conventional reprogramming delivers the reprogramming factors through viruses which retro-transcribe themselves into the genome. Because the integration is random, it has chances of causing insertional mutagenesis (as well as epigenetic changes in the nearby DNA). Furthermore, two of the reprogramming factors were oncogenes (cMyc being especially notorious), which has raised some fears as to the prospect of using these cells for therapy. Since then, alternative methods have avoided this issue[4], but seems as if at that time the authors, faced with this criticism, had wanted to defend the relevance of their work by repositioning the technology. In fact, the two applications go hand in hand, for developments of the technological platform are advancements for both its therapeutic and epistemic uses: fuelling any further research increased the resources for validating and improving the technology. In any case, the 'Disease in a dish' approach became a whole research paradigm, transforming the notions of diseases, patients, and variation, and integrating them in a particular epistemic sub-culture.[5]

[4] Technological developments in this direction have focused on non-integrating technologies. Episomal vectors, for instance, reside (and replicate) in the cell independently from chromosomes, thereby avoiding insertional mutagenesis. Certainly the cleanest (and most expensive) method, however, is the direct and transient delivery of the reprogramming factors in mRNA molecules (see for instance Warren et al. 2010). Finally, the very low efficiency of reprogramming increased the risk of artefacts by putting a strong selection pressure on the cells, something which is avoided by 'deterministic' reprogramming (Rais et al. 2013).

[5] For a review of iPSC disease modelling, see Grskovic et al. (2011), Cherry and Daley (2012). See Merkle and Eggan (2013) for a detailed discussion of an iPSC-based research pipeline. For iPSC models of neurological diseases, see Han et al. (2011). For an example of potential treatment, see Wu et al. (2011).

5.2.2 Repeated development

Because they were expected to replace embryonic stem cells (ESC) for many purposes, iPSC were compared to ESC from very early on[6]. The equal potency of ESC and iPSC was gradually established by accumulation of supportive observations. In their paper porting reprogramming to human cells, Yamanaka's lab demonstrate pluripotency by differentiating into neuronal cells and cardiomyocytes. The choice is not arbitrary, for these cell types have a considerable rhetorical power, partly because they are the cell types were iPSC technology promises to be relevant, but also because of the visual immediacy of the differentiated phenotype. In fact, the paper was even accompanied with a video: "twelve days after the induction of differentiation [into cardiomyocytes], clumps of cells started beating" (Takahashi et al. 2007, p.865). It is interesting that *in vitro* heartbeats played the same rhetorical role they did a century earlier in Montrose Burrows' pioneering tissue culture experiments (see Landecker 2007).[7]

Nevertheless, differentiation into one or two cell types is not sufficient to claim equal potency to ESC. In his time, Gurdon had demonstrated successful reprogramming by the formation of tadpoles, but in mammals cloning is considerably more difficult to achieve. Quickly, a balance was struck between feasibility and epistemic power, and still today the gold standard of pluripotency is the formation, upon implantation in immunodeficient mice, of teratomas (usually benign tumours) displaying tissues of the three germ layers (see also the discussion in appendix B). Cells able to differentiate into the three germ layers are by extrapolation expected to be able to form all the cell types of the organism[8]. As teratoma assays are expensive and time-consuming, they might in the near future be replaced by more cost-effective *in vitro* assays (Bock et al. 2011).

As some authors argued, "in technical and application terms, these two types of cell [ESC and iPSC] may be together considered as human pluripotent stem cells that differ in that the

[6] In fact, Hauskeller and Weber (2011) have argued that the need to compare iPSC to human ESC (hESC) prompted a stabilization of hESC, an entity whose nature had been somewhat unclear and debated.

[7] On visibility, see Nowotny and Testa (2011), as well as Landecker (2007) on the rhetorical power of movement.

[8] Nevertheless, some biologists remained sceptical, arguing that "it is not clear that iPSCs can display all the properties of ESCs: in mice where more stringent tests can be applied, murine iPSCs, unlike ESCs, have not yet formed live young in tetraploid aggregation assays." (Colman 2008, p.237) The demonstration was reported in 2009 (Boland et al. 2009, Zhao et al. 2009), after which the pluripotency of iPSC became relatively established.

former are 'embryonic' and the latter 'induced'." (Sasai et al. 2008, p.S52) Nevertheless, similarity beyond the mere differentiation capacity was a topic of heated discussions. Already in 2007, Rudolf Jaenisch's group at the MIT published in *Nature* a thorough transcriptomic and epigenetic comparison, concluding that "the biological potency and epigenetic state of in-vitro-reprogrammed induced pluripotent stem cells are indistinguishable from those of ES cells." (Wernig et al. 2007) However, the polemic persisted as alleged epigenetic differences were uncovered (see for instance Kim et al. 2010; see Yamanaka 2012 for a critical discussion).

From a certain point of view, it would seem that if iPSC and ESC have the same differentiation potential, other differences between them are irrelevant. Their relevance comes from the fact that iPSC were taken to be all-purposes surrogates of ESC, which required that they be *replicas* (see Chapter 3, section 3.3.2). Their equivalence served to establish the reliability of the derived cells. If the differentiated cells were derived from ESC-like cells, and if differentiation protocols "closely mimic mammalian development" (Chamberlain et al. 2008, p.227), then these cells were sure to be good models of their *in vivo* counterparts.

One can obviously validate the differentiated cells by comparing them to the target cells, however this approach is limited for two main reason. First, our access to the target cell type is often limited, and indeed the power of the reprogramming technology is to avoid this problem. Second, as I will discuss in the next section, cell types are generally not exhaustively defined, so that it is not always clear whether we are comparing the corresponding states. For this reason, scientists have generally added support to the validity of their model by demonstrating its similarity to developmental processes. In a recent review, two scientists of the Harvard Stem Cell Institute make this point in their distinction between directed differentiation, in which "the signaling pathways responsible for making the target cell type in vivo are stimulated", and 'direct programming' which accomplishes the same result without pretending to mimic development. Regarding the latter, they note that "it is still unclear to what extent these programmed cells are suitable for in vitro disease modeling, because they may be less similar to their in vivo counterparts than cells generated by directed differentiation." (Merkle and Eggan 2013, p.660)

This aim of mimicking development is not specific to iPSC, but is found in stem cell models more generally. It is most obvious when we look at authors' justification of their methods. For instance, Chamberlain et al. (2008) repeatedly establish a correspondence between development and the *in vitro*, especially with respect to timing, e.g. "The timeline of motoneuron development is also similar to in vivo development" (Chamberlain et al. 2008, p.232). Protocols – culture conditions and reagents added in a rigid timeline – effectively *model* the inner milieu of the developing organism, including its timing:

> "hESC aggregates are grown in suspension culture with hESC medium for 4 days. This step *approximates gastrulation* and induces the formation of the ectodermal germ layer." (Chamberlain et al. 2008, p.230, emphasis added)

In a similar way, Mariani and colleagues report the generation of "3D self-organized structures from human iPSCs that recapitulate the program of early dorsal telencephalic development in humans" (Mariani et al. 2012, p.6). By far the most astonishing such results has been the recent *in vitro* derivation of cerebral organoids (Lancaster et al. 2013), shown in Figure 5.1. Lancaster et al. have performed extensive immunohistochemical staining for markers of different brain regions, showing that the organoids "recapitulate various brain region identities" (Lancaster et al. 2013, p.2), starting with forebrain/midbrain/hindbrain but also including, for instance, different cortical regions (spontaneous neural activity was also detectable). As a proof of concept, the authors used patient-derived organoids to study microcephaly.

It is astonishing that the iPSC-derived cells are very often not said to model the organism's cells of interest, but to *give access* to them: "Using iPSC, researchers are beginning to have an idea of how to study developing neurons from patients" (Freitas et al. 2012, p.19). Indeed, iPSC are said not only to "model normal development", but to even allow "repeated development" (Nishikawa et al. 2008, p.727). Importantly, the idea of 'repeated development' implies not only an identity with the *in vivo* development, but something even more powerful, namely that this development is repeatable several times, transforming a unique organismal process into an experimentally and statistically tractable process.

What is particularly interesting here is that differentiation is at the same time an object of study and a means of producing study materials. Differentiation is perhaps primarily a means of producing differentiated cell lines which can then be compared across conditions.

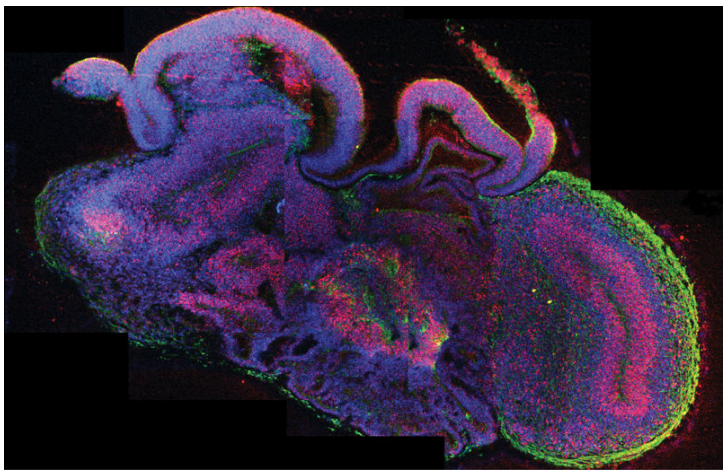

Figure 5.1: Stained section of a three-dimensional 'cerebral organoid' derived *in vitro* from human pluripotent stem cells (adapted from Lancaster et al. 2013, Figure 1-C). Cerebral organoids can grow up to 4mm in diameter, and develop heterogeneous regions reminiscent of early brain regionalization. Immunohistochemistry here shows nuclei in blue, SOX2-positive cells (neural progenitors) in red, and TUJ1-positive cells (neurons) in green.

At the same time, differentiation is part of development, and can therefore be in itself the focus of research. For instance, Pasca et al. (2011) used the cellular reprogramming technology to study Timothy syndrome (a genetic disease associated with developmental delay and features of autism spectrum disorders). Upon neuronal differentiation, they noticed that cells from the patient differentiated more towards upper layers of the cortex than the cells from the controls. At this point, one could have sorted cells and isolated comparable subpopulations in the two conditions, and studied these cells. Differentiation would then be taken solely as a means of producing the desired cell type. However, the skew observed in the differentiation is in itself indicative of a bias in differentiation potential which is potentially disease-relevant. In this context, the differentiation protocol is both a means of producing cortical neurons, and of measuring differences between the conditions.

Not only differentiation, but the very act of reprogramming can also in itself be a detection tool, and in Chapter 3 I have given an example of this in cancer research (section 3.2.2). Even failure of reprogramming is informative. For instance, it was observed that fibroblasts and keratinocytes of patients suffering from Fragile X syndrome (an inherited mental retardation syndrome caused by repeat expansions in the FMR1 gene) could not be reprogrammed unless the mutation was first corrected (Urbach et al. 2010). The fact that the mutated gene is involved in reprogramming is potentially informative of both the disease and the regulation of pluripotency.

Differentiation, reprogramming, and transdifferentation all oscillate between being epistemic things and technical tools, and even as technical tools, they can be production procedures or detection procedures (see section 3.3.1). This creates research opportunities, but it creates problems as well. When comparing cells from two biological conditions (e.g. diseased and control) which have undergone differentiation, a major difficulty is to make sure that the cells are in fact comparable – that they compare equivalent cell-states in the disease and in the control. Suppose that we differentiate iPSC cells from patients and controls into neural progenitors, and notice a difference (in, say, gene expression profile, or morphology). Should this be taken to mean that patient-derived neural progenitors are different from controls, or could it be that the 'diseased cells' react differently to the differentiation protocol, so that in effect we are not comparing cells at the same differentiation stage? Answering this question requires a notion of what it means to be a neural progenitor.

5.2.3 Cell types between in vivo and in vitro

It is common to attribute an essentially (*in vivo*) positional identity to cells reprogrammed and redifferentiated *in vitro*. It is for instance argued that "[i]f we hope to use [induced-neurons] to model [Parkinson's Disease], it is imperative that these neurons are genuine human midbrain nigral neurons." (Drouin-Ouellet and Barker 2012, p.2). That a skin fibroblast can become a 'genuine' midbrain cell is the magic of molecular cell biology, which managed, through the constitution of cell types, to internalize in the cell much of its relation to the rest of the body.

In fact, however, the construction of cell types is a problematic issue that has been extensively discussed by scientists. A most remarkable example is the review published in *Nature* by Luciano Conti and Elena Cattaneo (Conti and Cattaneo 2010). They note that "[t]he identification of early stage-specific neural markers has allowed neural induction to be followed both *in vivo* and *in vitro*." (Conti and Cattaneo 2010, p.176) Because markers (typically cell surface molecules) can be tracked *in vivo* and *in vitro*, they allow the establishment of corresponding cell states in each environment. Figure 5.2, taken from the paper, shows correspondences between *in vivo* developmental stages and cell populations characterized *in vitro*. The process of internalization of positional cell types – or, conversely,

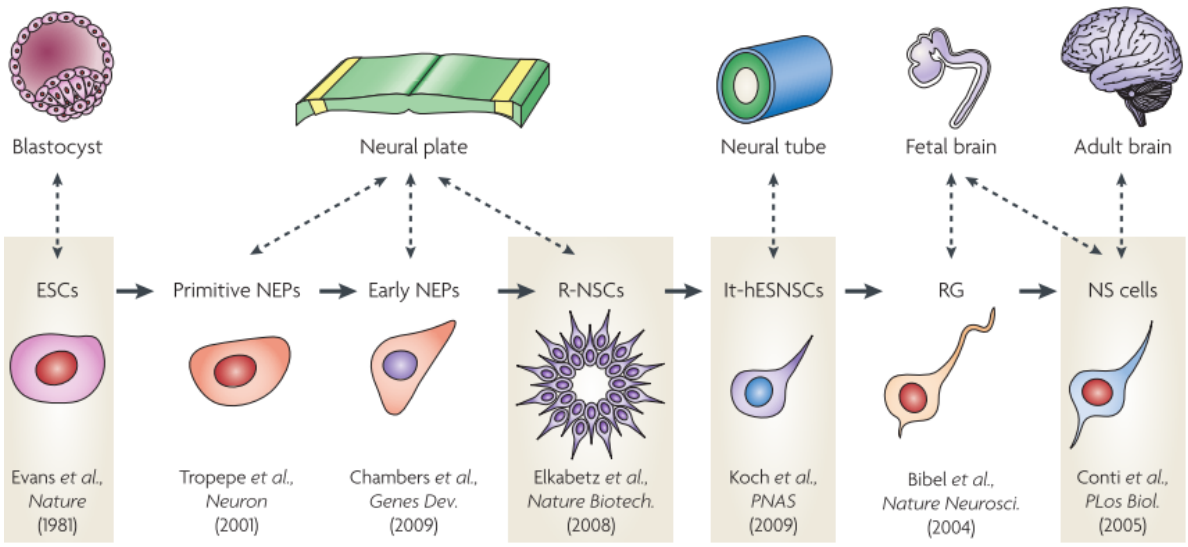

Figure 5.2: Correspondence between *in vivo* developmental stages and populations characterized *in vitro*, adapted from Conti and Cattaneo (2010).[11]

of the spatio-temporal location of structurally-defined cell types – is obvious from several passages of the paper. For instance, it is written that in early mammalian brain the neural tube is "made up by a layer of so-called neuroepithelial progenitors (NEPs)" (Conti and Cattaneo 2010, p.176), which however can also be obtained from hESC through *in vitro* manipulations[9]. It is important to note that the construction of a cell type can come from either, or most often from both, the *in vitro* or the *in vivo*. For instance, the authors note that a given cell state obtained and characterized *in vitro* could "possibly embod[y] an intermediate developmental stage between rosette-organized NEPs and [radial glial cells]" (Conti and Cattaneo 2010, p.178), because it possesses some properties of both NEPs and radial glial cells. Most often, the construction starts from the *in vivo*, or from a back-and-forth between the two.

What I would like to emphasize here is that (as indicated by the arrows in figure 5.2) the establishment of the correspondence does not solely rely on the correspondence in marker

[9] More specifically, through "a dual-SMAD inhibition protocol for rapid (6 days) neuralization of hESCs, generating a population of 'early' SOX1-, PAX6-, OTX2- and FOXG1-expressing NEPs." (Conti and Cattaneo 2010, p.177)

[11] Original legend: "Developmental links between the different NSC populations that can be isolated or generated *in vitro*. Shaded boxes indicate the neural stem cell (NSC) populations that can be obtained through mid-term or long-term expansion *in vitro*. The corresponding in vivo developmental stages and the reports that first described these populations are indicated. Induced pluripotent stem cells, generated by means of reprogramming, are thought to have the same developmental potential as embryonic stem cells (ESCs). Late NSCs have not been included as they have many similarities (in terms of growth factor requirements, antigenic profile and neuronal subtype specification) with NS cells, probably indicating that they represent an analogous population grown in different ways (monolayer and aggregation)." (Conti and Cattaneo 2010, p.178)

expression between one step of the process, but also of the more general correspondence between the two developmental processes: the sequence of *in vivo* development (above) seems to be replicated *in vitro* (below). In other words, the correspondence between, say, cells of the neural plate and early NEPs does not solely depend on their similarity, but also on their corresponding position within an overarching developmental structure. In other words, our confidence that NEPs model cells of the neural plate does not depend solely on the relations between the two, but also depends on the relation of NEPs with other cellular models, and of the relation of these with the target system. This is an important point which I will later try to capture under the label of distributed modeling.

The *in vivo-in vitro* correspondence is not unproblematic, and indeed the authors discuss some important epistemological issues related to early neural cell states, especially neural stem cells (NSC).

> "Neural stem cells (NSCs) are self-renewing and multipotent populations present in the developing and adult mammalian CNS. They generate the neurons and glia of the developing brain and also account for the limited regenerative potential of the adult brain." (Conti and Cattaneo 2010, p.176)

NSCs are initially defined as a developmental phenomenon *in vivo*, identified for instance in the adult subventricular zone (SVZ) and hippocampus. The cells have some properties (self-renewal and multipotency) and functions within development. Because these properties can be detached from development (as a spatio-temporal process), they provide an identity criterion that allows the cells to be tamed *in vitro*. Interestingly, the authors consider these properties to be directly operational: NSCs are "cells that are operationally characterized by self-renewing and multipotential differentiation" (Conti and Cattaneo 2010, p.176) Later on, the authors explain that NSCs can be purified and expanded through the neurosphere assay:

> "In these [low-adherance] conditions, most differentiating or differentiated cells are expected to die [due to anoikis], whereas the NSCs respond to the mitogens, divide and form floating aggregates (primary neurospheres) that can be dissociated and replated to generate secondary neurospheres." (Conti and Cattaneo 2010, p.181)

The neurospheres contain different cell types, and therefore the assay ensures that the neurosphere-forming cells are both multipotent, and capable of self-renewal. Because it is

linked to a protocol, the notion of neurosphere-forming cells is operational, and although it is *an* operationalization of the self-renewal and multipotent features of the cells, it is by no means the only one. NSCs self-renew and are multipotent *in vivo*, but in very different conditions – in fact the authors note the discrepancy between neurospheres and the *in vivo* NSC niche: "neurospheres do not show any cellular organization in terms of cell types and distribution that may recapitulate the SVZ structure." (Conti and Cattaneo 2010, p.181) The neurosphere assay is an operationalization of self-renewal and multipotency, but detection of the same properties *in vivo* in the SVZ both requires and represents a different operationalization. Whether the two match remains an open question, which the authors acknowledge:

> "The data described above call for extreme caution when extrapolating *in vitro* results to normal development or physiology without corresponding *in vivo* data and suggest that the self-renewal and multipotency demonstrated by NSCs *in vitro* might result from exposure to growth factors that create a synthetic transcriptional and biological state." (Conti and Cattaneo 2010, p.184)

The authors thus present NSCs as either an *in vitro* artefact or an *in vivo* reality. Another avenue, suggested by the discussion of the previous chapter, would be to consider cell types as theoretical terms. In any case, it is important to point that a full correspondence is not necessary between the two. Correspondence, and in fact the rhetorics of mimicking, giving access to, or repeating development is square into the aim of obtaining replica, in the sense described in Chapter 3. Measuring devices, in contrast, do not require such a mimicking, and their calibration proceeds differently (see section 5.3.3). Nevertheless, the assumption of correspondence is a productive heuristic (see section 4.3.2), and as I will argue in the next section, the productivity of the *in vitro-in vivo* lies precisely where they do not match.

5.2.4 Turning the inside out and recomposing it

The oscillation between reproduction and contrast has been characteristic of the concept of *in vitro* since its very beginning and throughout its changes of meaning. In the 19th century, *in vitro* depicted the absence of life, and Eduard Buchner's feat (Buchner 1897) was to produce a reaction specific to life (fermentation) in the absence of life – i.e. in the absence of yeast cells. Likewise, in his history of the discovery of the mechanisms for protein

synthesis, Rheinberger (1997) uses the expression *in vitro* to designate the absence of whole cells. As Rheinberger writes, "the prevailing momentum of early molecular biology resided in creating the technical means of an extracellular representation of intracellular configurations." (Rheinberger 2000, p.19) Nowadays, however, to most biomedical researchers (and unless otherwise specified) *in vitro* just means in a dish – as opposed to in an organism, but with whole cells. In other words, cells in a Petri dish can be considered, depending on the context, either *in vivo* or *in vitro*[12]. This ambiguity in meaning is due to the fact that the *in vivo* / *in vitro* distinction maps a relative, rather than fixed boundary.

The *in vivo* of protein synthesis is the cell, because protein synthesis is a phenomenon normally occurring inside cells. Instead, *in vitro* systems are merely ways of probing, studying, and manipulating this phenomenon. The same can be said regarding the other referent of *in vitro*: in studying development – an organism-level phenomenon – cell culture becomes a technical means of an extraorganismal representation of intraorganismal configurations. *In vitro*, therefore, does not refer to the presence or absence of cells, but rather to an isolation of parts of a higher-level phenomenon[13].

Another classical example of *in vitro* research is the work of the 1912 Nobel prize laureate Alexis Carrel regarding the immortality of somatic cells (where immortality is to be understood in terms of lack of ageing – for an interesting discussion of the concept of immortality, see Hayflick 2000). Following a debate in theoretical biology about whether ageing was a cellular or an organismal phenomenon, Carrel (1929) attempted to solve the issue experimentally (see Landecker 2007). By providing a culture environment where waste was evacuated and the cells were provided with fresh nutrients, Carrel showed that the cells did not stop growing, and claimed to have maintained them for decades. It is interesting that just like Carrel himself dismissed earlier results as due to the inability of the experimenters to adequately cultivate cells, his results were also later dismissed as an experimental error. In light of Leonard Hayflick's discovery of senescence (Hayflick 1965)

[12] In a similar way, Rheinberger notes that "[w]hether an entity is considered natural or artificial depends on what one is doing with that entity: If one works with an in vitro system, every whole cell therein behaves as an artifact." (Rheinberger 1997, p.198)

[13] The notion of levels of organization is ubiquitous but generally undefined. Given my coherentist conception of science (see especially section 4.4), it is sufficient for me that science represents entities and phenomena as organized into levels. However, if one were to seek an explication of the notion of level, Wimsatt's treatment of the issue appears to me as the most convincing. The gist of his proposal is that "[l]evels of organization can be thought of as local maxima of regularity and predictability in the phase space of alternative modes of organization of matter." (Wimsatt 2007, p.209)

– that cells stop dividing after a certain number of divisions, due to the shortening of their telomeres – the consensus view is that Carrel's medium was most likely contaminated with fresh cells. In a twisted way, however, Carrel was right: had Hayflick cultivated his cells 'in the right conditions', they could have replicated indefinitely, although they would not have remained 'the same cells'. Thanks to cellular reprogramming, we know that virtually any somatic cell, in the right culture conditions, has the potential for infinite replication – that at least some of the hallmarks of physiological time can be reversed. Together, these findings are at the same time a solution to and a dissolution of the initial problem: answering the Carrel-Hayflick debate would necessitate an agreement about the contentious 'normal environment' in which cells should be said capable of infinite proliferation. Instead of offering a yes or no answer, *in vitro* experimentation did much more, and allowed to identify the factors determining self-renewal capacity.

Gradually, biologists recompose, in a dish, the cells' *in vivo* environment. This is sometimes out of practical necessity: in the first decades of tissue culture, cells would not survive in culture without extracts from the organism, and even today serums are still widely used for cell culture[14]. Often, however, it is for modeling purposes, and the cerebral organoids discussed earlier (see Figure 5.1) provide an amazing example. The aim, however, is never a full recomposition: the very nature of an *in vitro* system is precisely the selectivity of its environmental modeling. Would the *in vitro* model every aspect of the *in vivo*, it would cease to be useful.

The kind of purification intrinsic to *in vitro* systems is continuous with other ways of experimenting, and this most obvious in transplantation experiments. Rheinberger (2010) discusses an interesting example to this effect, namely the work of Ernst Caspari in Alfred Kühn's lab in the 1930's on the pleiotropic effect of the red-eyed moth mutant. He could show that the color phenotype of the mutant was mediated through some intermediary substance by transplanting of the testicles of a mutant into a wild-type. "Thanks to the transplantation technique," writes Rheinberger, "their system allowed direct access to physiology." (Rheinberger 2010, p.104) However, the key point here is that transplantation

[14] Serum is however becoming less and less popular, not because of its yield, but for reasons of standardization: its content is largely undefined, and can vary from batch to batch. For the value of the *in vitro* lies in control, and its improvement is not the maximal inclusion of more and more from the *in vivo*, but the increase of scientists' ability to selectively do so in a controlled way.

allowed the causal isolation of the effect of the mutation in the testis from that in the rest of the organism. In a way, transplantation can be thought of as a method to randomize the whole organism save for one organ. It is one step in a progressive effort at isolating causal influences: a couple of years later, the project moved "from the transplantation experiment with living tissues to a chemical-physiological attempt to obtain extracts from the blood and organs as well as the incorporation of organ extracts." (Westphal, quoted in Rheinberger 2010, p.113)

In this light, much of *in vitro* research should be understood not as explanation, but as 'transplantation in a dish', for purposes of isolating causal influences. That was noted by Weber, who writes that "the great epistemic value of reconstitution experiments is due to the possibilities for controlled experimentation that they offer" (Weber 2005, p.146) One of the famous findings of iPSC models provides a good example to this point. In the study of iPSC-based models of amyotrophic lateral sclerosis (a motor neuron debilitating disease, also known as Lou Gehrig's disease), it was shown that neuronal death was not an autonomous phenomenon of the neurons themselves, but was instead due to toxicity of the surrounding mutant glial cells (Marchetto et al. 2008). Such a 'non-cell-autonomous'[15] effect can only become visible if one has the capacity to grow neurons both alone and in co-culture with glial cells. This would not have been possible *in vivo*, but depended on an incomplete, or differential, recomposition. It is the sheer amount of such contrasts that can potentially be made that make *in vitro* systems so powerful:

> "these studies confer the ability to test various neuron non-cell-autonomous effects, such as inflammation, oxidative stress, activity-dependent modulations and the influence of stress hormones in psychiatric disorders." (Brennand et al. 2012, p.11)

This departs from the common assumption that *in vitro* approaches are *per se* reductionistic[16]. For one thing, scientists never assume, nor do their approaches require, that the *in vivo* phenomenon will be *entirely* reproducible in a dish: for an *in vitro* approach to

[15] Cell-autonomous phenomena are those that can be observed in single cells, while non-cell-autonomous phenomena result from the interplay between cells.

[16] Such an assumption is also held by scientists practicing such a science: "Thus, the very fact that the cells are isolated means that we lose any information concerning adhesion proteins, selectine, cadherin and others, which keep the cells associated in an organ, as well as any information about lacunary and communicating intercellular junctions that act as channels through which ion flow occurs." (Vignais and Vignais 2010, p.221)

be successful, it is sufficient that a part of the phenomenon will be reproducible. More importantly, the *in vitro* is always studied in comparison with the *in vivo*, and it is precisely the contrast between the two that is productive. In fact, experiments such as those of Carrel precisely depend on a discrepancy between the *in vitro* and the *in vivo*. It is this contrast itself that allows the the identification of *explanans*. As Rheinberger writes, "[o]ne of the most important procedures for producing resonance in the life sciences is the mutual superposition of in vivo and in vitro approaches." (Rheinberger 1997, p.65)

5.3 The iPSC research paradigm

5.3.1 iPSC models and human variation

Many of the perks of iPSC models are shared with other stem cell models, and in fact there have been successful examples of ESC-derived models of diseases, most prominently from discarded embryos following screening for genetic diseases (see for instance Biancotti et al. 2010) or by genetically engineering healthy stem cells (for instance Pickering et al. 2005 for cystic fibrosis). The former approach is obviously limited in yield and can only work for very common disorders, while genetic engineering was until recently a dirty and time-consuming process. However, it has been argued that the development of contemporary technologies, such as Transcription activator-like effector nucleases (TALEN), shifted the balance. Ding et al. (2012) have for instance argued that the development of engineered lines was even faster than deriving iPSC from patients, provided that the time needed to recruit patients is taken into account. Genetic engineering also has the advantage that except for the inserted mutations, diseased and control lines are completely isogenic. Unless they also rely on genetic engineering, researchers using iPSC models have to compare cells from patients with, in the best of cases, cells from unaffected relatives. Of course, TALEN-editing is possible only for monogenic (or simple) diseases, and moreover for those cases where the responsible loci are already known. Part of the value of iPSC is the ability to model a disease – or even disease susceptibility – even without knowing its precise cause.

More importantly, iPSC "are made from living donors with detailed medical histories attesting to the impact of the disease on particular individuals" (Colman 2008, p.237).

Because they are relatively easy to derive, it is possible to study a disease over a variety of genetic backgrounds, thus leading to results that will be more robust across different patients. Note that this advantage is necessarily opposed, and contradictory to the isogenic aspect of genetically engineered lines: greater control comes at the cost of lesser representativity. This dilemma is by no means new. After the pioneering work of Clarence C. Little in the genetic standardization of laboratory mice in the first half of the 20th century (see Rader 2004), the next step was to push the standardization of the environment up to the animal's innards. For instance, at the fourth symposium of the International Committee on Laboratory Animals held in 1969, the famous geneticist George W. Beadle (1958 Nobel prize laureate with E.L. Tatum) urged to pursue standardization to germ-free animals[17]. However, in the same meeting, scientists of the Laboratory Animal Institute in Hungary were instead suggesting to work on mosaic populations to avoid strain-specific artefacts (International Committee on Laboratory Animals 1971, p.218-229). Their point was not to abandon inbreeding, for mosaic populations needed to be made in reproducible ways, crossing inbred lines in a controlled way. Nevertheless, any variation introduced in the test population will inevitably reduce the statistical power of the test, a point which was raised (by Thomas B. Clarkson) in the meeting's discussion: "One of the main aims of experimental design is to reduce within-group variation. This increases precision. Don't mosaic populations reduce precision?" (International Committee on Laboratory Animals 1971, p.229)

Insofar as they are construed as replicas of (cells of) heterogeneous human patients, iPSC meet this dilemma between experimental tractability and representativity. However, if we cease to consider the model as a surrogate for patients, this variation may itself become a tool enhancing experimental tractability. Using diverse genetic backgrounds allows the effect of a specific genetic variation to stand out, screening out patient particularities. This is more than merely enhancing the discriminative power of assays, and can also be harnessed to offer new insights into pathology. Consider the following passage:

> "In polygenic and complex diseases, ignorance regarding gene loci should not be an obstacle to an iPSC approach – it may even facilitate their identification.

[17] "Unless germfree animals are used, which is expensive in both energy and dollars, the laboratory animal researcher must work with ecological systems that are difficult to keep constant." (International Committee on Laboratory Animals 1971, p.155) On germ-free animals and standardization, see Kirk (2012).

> Syndromes, defined as diseases with several distinct clinical features, are of particular interest when patterns of aberrant gene expression and pathology cannot be superimposed, because this discrepancy may point to compensatory mechanisms of potential therapeutic interest occurring in the unaffected tissue." (Colman and Dreesen 2009, p.245)

Complex diseases are intractable both due to the heterogeneity of their causes and to the complexity of the phenotypes. Because iPSC offer different instantiations of the disease on different genetic backgrounds, their intersection allows the identification of a subset of the clinical features with a subset of the underlying causes[18]. At the same time, what does not intersect – either among clinical features or the underlying causes – stands out as targets of explanation, thereby offering an entry-point into the mechanisms regulating these features. This point is not new to iPSC, and much of Genome-Wide Association Studies (GWAS) follows this logic. Consider, for instance, the GWAS studies on hypertension, notably those performed in the context of the Wellcome Trust Case Control Consortium. An important study counting tens of thousands of participants identified 8 genetic loci associated with blood pressure (Newton-Cheh et al. 2009), but together they explained less than 1% of the variation in blood pressure within the population. The number of associations found was expected to increase linearly with sample size, and indeed it did (Ehret et al. 2011). Given the very low proportion of variance these loci explain, it is not of much clinical relevance to a given individual whether he has these variants or not. Indeed, the main value of such associations is not to predict the risk of individuals, but instead to provide entry-points into disease mechanisms.

The situation suggested by the above quotation from Colman and Dreesen adds another dimension to this. In the case of the hypertension GWAS, one expects that the association loci lead to the identification of players in the pathways influencing the disease of interest, each of the patients instantiating the influence of these pathways. In the iPSC situation suggested by Colman and Dreesen, the patient is not anymore the target of the modeling relationship: the model, understood as the set of cell lines derived from the whole cohort of patients, represents something which strictly speaking is not instantiated in any of patients.

[18] The strategy is in line with the recent trend in nosology, especially in the context of research. In psychiatrics, the Research Domain Criteria (RDoC) of the National Institute of Mental Health proposed to replace symptoms-based categories with a dimensional system which "places equal weight on behavioral functions and upon neural circuits and their constituent elements" (Cuthbert and Insel 2013, p.6). A similar trend can be observed in other fields, as exemplified by the Committee on a Framework for Development a New Taxonomy of Disease (2011).

Instead of instantiating or representing the disease, the patients are themselves models – even *partial* models – in that they are indirect means of accessing it. If they represent the disease, it is in that they are different spaces of representations in which it is projected. This point was raised by Rheinberger in the context of model organisms:

> "Under the epistemic regime of the early twentieth century, biological differences between research organisms started to be transformed into tools that could be exploited for the characterization of the most general features of living beings. [...] A model organism, then, is no longer analyzed in its own right: it is investigated for the sake of something that lies beyond it." (Rheinberger 2011b, p.163)

In such contexts, it is not anymore patients which are the target of modeling, but a more abstract disease entity, and this intermediate between models and patients allows for more elaborate modeling relationships.

5.3.2 In vitro symptoms as intermediary anchorpoints

Figure 5.3, adapted from Bellin and Marchetto (2012), is a good representative of a variety of similar figures illustrating the iPSC 'research pipeline'. Somatic cells are harvested from a patient through a biopsy, before being reprogrammed to iPSC and differentiated to the relevant tissues. The derived cells can then be used for every step of drug development. They first serve as disease models yielding insights into the pathogenic mechanisms and, hopefully, druggable targets. Potential drugs can then be tested on the same cells to see if they correct the *in vitro* disease phenotype. Finally, the iPSC can also be used to generate specialized cells for toxicity testings. This is what warrants the label of "Human preclinical trials 'in a tube' ".

Bellin and Marchetto are very provocative in their usage of medically loaded concepts in the context of the dish, much beyond the simple expression 'diseases-in-a-dish'. They write of "studies for treating diseases 'in a dish'," of "the symptoms in human iPS cells and their derivatives" (Bellin and Marchetto 2012, p.713), and proposing iPSC as "the new patient" (Bellin and Marchetto 2012, p.724) The rhetorics of surrogacy hide some of the complexities of this process, which I will analyze in the next section. But they also highlight a central point of the iPSC-based research paradigm, namely the critical importance of a disease-relevant phenotype *in vitro*: in order to screen for drugs, there has to be an *in vitro*

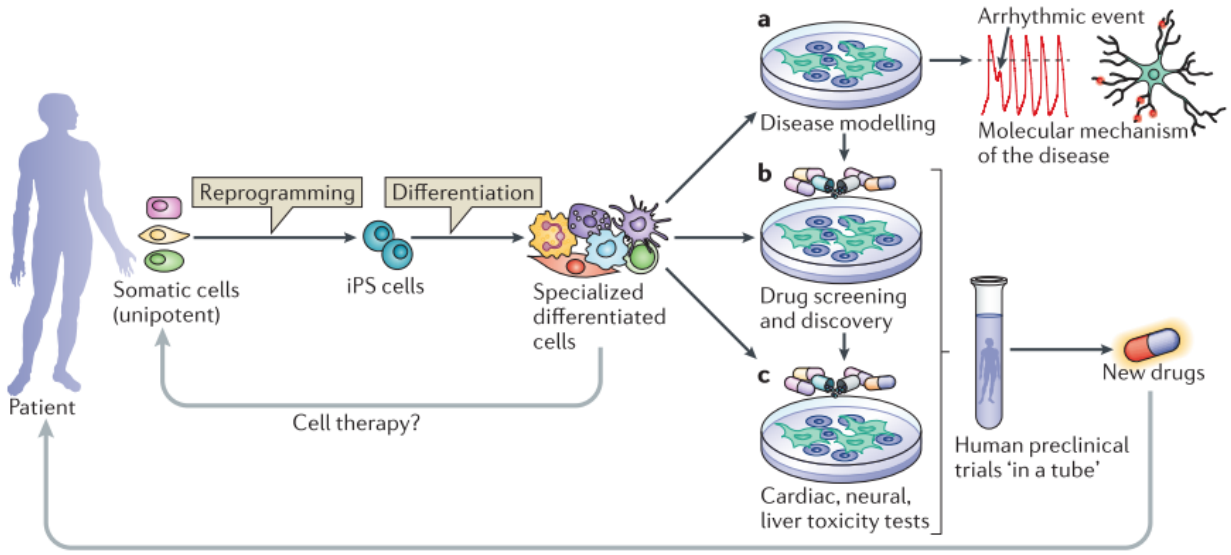

Figure 5.3: The iPSC research platform. Adapted from Bellin and Marchetto (2012).

'symptom' that the drug is expected to relieve, and even for a mechanistic understanding of the disease there has to be something *in vitro* which is dissected. For this reason, *in vitro* phenotypes have been a primary concern of the iPSC disease modeling community.

There is, here, an interesting analogy to be made with an important episode in the history of molecular biology. In an insightful contribution on "The transformation of Molecular Biology on contact with higher organisms", Morange (1997) made some very interesting remarks regarding what has come to be known as the 'crisis of molecular biology'. After their striking success in understanding the biology of simple organisms such as *E. Coli*, molecular biologists felt confident that they could tackle a molecularization of developmental biology, but major difficulties in doing so led to a crisis in the field. A central element of Morange's diagnosis is the following:

> "The roots of the crisis should be sought at the epistemological level: what molecular biologists cruelly lacked, what led them to a feeling of decadence, was the total absence of a definition of ... what would be an explanation of development. What was missing was the distinction between the 'explanans', which was thought to be molecular, and the 'explanandum'. Molecular biologists were eager to conceive what might be a molecular explanation of development, not a molecular description of it. The urge was not the same for them, who were in search of a causal explanation of development, and for 'classical' biologists and embryologists for whom, by tradition, an extensive and complex description was a sufficient goal." (Morange 1997, p.387)

The insistence on explanation rather than description was driven by molecular biologists' urge for universal explanations, which embryology was not providing. However, this shift

> "required that another level of description of the biological facts not be discovered, but valorized. This level was the cellular level, and this explains the dramatic development of cell biology during these years. Cell biology provided what Harold Kincaid called the 'place holders', the terms which are introduced to designate an entity, a process for which we have good evidence, but whose precise nature is unknown." (Morange 1997, p.390; the reference is to Kincaid 1990)

The phenomena of cell biology played the role of *explananda* for which molecular biology could provide an explanation. In many ways, iPSC-based models offer the same solution to molecular medicine when faced with the challenge of disease. This is why most of the modeling effort is aimed at identifying differences, visible in the dish, between control- and patient-derived cells – or, as they are usually designated, between healthy and diseased cells. Such differences constitute an *in vitro* phenotype which can be used for several purposes, such as providing an endpoint to be assessed in the context of *in vitro* drug screenings. More fundamentally, the importance of this phenotype is analogous to the contribution Morange described for cell biology: the *in vitro* phenotypes provide a necessary intermediate between molecules on the one hand, and development or pathology on the other. This intermediate can then become the *explanadum* for a molecular explanation – and, potentially, the target of a molecular intervention. This is especially relevant in the context of neurodevelopmental diseases, where there is a huge gap between genetic difference-makers and the clinical phenotype. With *in vitro* phenotypes, scientists are linking the two neither from a bottom-up or a top-down approach, as has traditionally been the case in biology, but are attempting to bridge the gap by finding an anchorpoint (sometimes shooting in the dark) somewhere in the middle.

5.3.3 Translating phenotypes

Careful reprogramming (avoiding genetic modifications of the cells and controlling for all potential errors) is not simple, and in many cases the directed differentiation remains a major challenge (see Unternaehrer and Daley 2011, p.2283; Grskovic et al. 2011, p.8). Nevertheless, because the basic cell reprogramming technology is both cheap and easy, it has prompted hundreds of research projects. But as is clear from the pipeline presented in the previous section, research applications depend on the identification of an *in vitro*

phenotype, and scientists quickly realized that such a phenotype would not be so trivial to obtain. After all most patients, if they are to be patients at all, develop more or less normally, with all their organs, and their diseases are most often problems of fine-tuning. We should not expect major differences in a dish, the thought goes, but instead minor differences adding up in the context of the organism. In a 2011 review of the field, it was noted that out of 25 modeling studies cited, only 7 revealed a specific functional deficit (Unternaehrer and Daley 2011, p.2283). Scientists therefore began to develop search strategies to uncover *in vitro* disease-related phenotypes.

Ideally, the *in vitro* phenotype would be "the same as" (Marchetto and Gage 2012, p.642), or would "recapitulate" (Grskovic et al. 2011, p.1), the clinical condition. This is seldom the case, however, for cells generally lack most of the defining features of diseases (with the exception of some simple metabolic deficiencies). Even when it seems to do recapitulate the disease, the *in vitro* phenotype is always a translation of the clinical to the *in vitro* context: "the in vitro phenotype is a single cell-based extrapolation of a known in vivo phenotype" (Hinson et al. 2012, p.5). As a consequence, the *in vitro* phenotype could be very different from the *in vivo* one, for the disease might manifest itself in very different ways in these two micro-environments. This means that one should abandon the idea of surrogates (see section 2.4.1) – of mimicking the clinical. It also means, as I argued in section 2.4.3, that extrapolation is not only happening from model to modeled (from bench to bedside), but equally in the other direction:

> "Expected phenotypes based on previously established animal and cellular models and observations from neuropathological studies should serve as a means to establish hypotheses or help validate the specific iPS model but the identification of novel mechanisms or cellular phenotypes remains an exciting possibility." (Han et al. 2011, p.639)

Obviously, this is possible only when much is already known about the disease. In the case of Alzheimer's disease, for instance, the *in vitro* phenotype was expected from *post mortem* studies on patients, in which amyloid-beta plaques and tau tangles had been associated to the disease. Scientists immediately went to look for them, and found them in the dish (Israel et al. 2012, Shi et al. 2012). When such knowledge is poor, however, one has to go fishing for differences between patient- and control-derived lines.

In any case, *in vitro* phenotypes are not given, but are intensely sought for in a proliferation of assays. Every phenotype identified becomes a potential assay for other diseases. A first strategy, meant to detect phenotypes which are not cell-autonomous, was to grow the cells together with other cells types, namely those that are normally neighboring the neurons in the site of pathogenesis in the body. As mentioned earlier, this has for instance led to the discovery of the involvement of glial cells in amyotrophic lateral sclerosis. Another such strategy is to use 'stressors' (Mattis and Svendsen 2011, p.386): perhaps the cells behave in a normal way in most conditions, and display a functional deficit in particularly rough conditions – in the presence of waste, etc – which can be modeled *in vitro*. Even time can be modeled, or 'accelerated', which is of particularly relevant for late-onset diseases:

> "the sweeping term 'late onset' refers to the appearance of clinical symptoms; subclinical developments may occur a lot earlier and may be captured by the in vitro methodology. [...] Furthermore, emergence of a disease phenotype might be facilitated during in vitro culture resulting from suboptimal media formulations or by deliberately perturbing the system in a stress-inducing way (e.g., serum starvation, O2 reduction, heat shock, etc.)." (Colman and Dreesen 2009, p.245)

This approach was successful notably with iPSC from Parkinson's disease patients carrying a known mutation of the LRRK2 gene: upon exposure to a variety of oxidative stresses (e.g. hydrogen peroxide), mutant neurons died twice as much as control neurons (Nguyen et al. 2011).

The artificiality of these phenotypes however points to a risk of their being artefacts. Clearly, not anything will go, for phenotypes found – or indeed constructed – in the dish might not even have an *in vivo* correlate, or they might be side-effects that are irrelevant to the disease. Ideally, the pathological relevance of the *in vitro* phenotype should then be tested *in vivo*, typically in animal models of the disease. Once more, this requires a translation of phenotypes: both the (human) clinical phenotype and the iPSC-derived *in vitro* phenotype have to be translated to the animal model. The ideal corroboration, then, is to correct the *in vitro* phenotype in the animal model, and observe a change in the clinical phenotype – or to be precise in the animal extrapolation of the clinical phenotype. If the latter is attenuated, we have conclusive evidence that the phenotype is relevant for the disease. Note here that just as the animal model can be said to *model* the clinical phenotype,

it also models the *in vitro* phenotype, and this introduces an interesting complexity to the very notion of modeling.

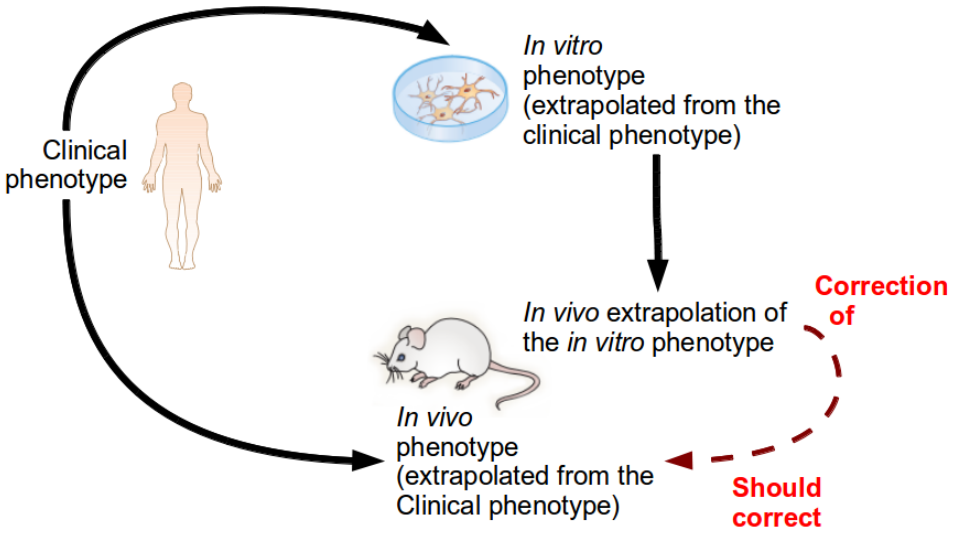

Figure 5.4: Translation and validation of disease phenotypes.

Figure 5.4 represents this functional validation, as well as the relations between the different phenotypes involved. The inappropriateness of the dyadic view of modeling (section 2.4) should be immediately obvious: to start with, this example of modeling involves at least three inter-related systems (the patient or clinical system, the *in vitro* system, and the animal model), of which at least two are models. More importantly, an account structured on a model-target dyad leaves unanalyzed the relations between the two models. For what makes the dish a good model for the patient depends on what is established in the third system.

An important thing to note, here, is that this validation in the third system makes it unnecessary for the *in vitro* model, for instance the cell type used, to be the same as in the *in vivo*: what matters is whether the *in vitro* phenotype is a good measurement. This requires that the *in vitro* phenotype has an *in vivo* correlate, which however needs not be a replica of the *in vivo*. Here we see again two modeling strategies reflecting the distinction made in Chapter 3 (section 3.3.2). The first, relying on an *in vitro* mimicking of development (section 5.2.2), produces *models-as-replica*. The second produces *models-as-instruments* – measuring devices – which rely not so much on a similarity relationship with the target system, but on connections with the rest of the research system, both its theoretical and its experimental components.

5.3.4 Comparative cellular models

Along with other members[19] of Giuseppe Testa's group (European Institute of Oncology), I am involved in the iPSC-based modeling of two neurodevelopmental diseases caused by symmetrical copy number variations of the 7q11.23 region, which encompasses about two dozens of genes (most of which are largely unexplored). Although multisystemic, the diseases are characterized most importantly by their cognitive-behavioral phenotypes. Interestingly, some aspects of the clinical phenotypes are symmetrical, reflecting the symmetry of the underlying genetic lesion. Patients who carry the hemi-zygous deletion of the region are affected by Williams-Beuren Syndrome (WBS), which is characterized at the cognitive-behavioral level by hypersociability and a clear pattern of cognitive impairment, with very poor visuospatial processing and comparatively preserved language skills (Bellugi et al. 2000). In contrast, duplication of the same region (7Dup) is associated with features of autism spectrum disorders, including severe language impairment and asocial behavior (Ferrero et al. 2010). Some other features, instead of being roughly symmetrical, appear to be shared between the two diseases: for instance anxiety and dysmorphic facial features.

The fact that the genetic symmetry is reflected up to the characteristic components of the clinical phenotypes is a strong indication that it will also be observable at intermediate levels of biological organization, or pathogenesis. Studying the two diseases together can therefore allow to distinguish those *in vitro* intermediate phenotypes that are also symmetrical, and therefore that are most likely to be relevant to the symmetrical features of the diseases. This is a first way in which these models question the dyadic view, especially the independence of different models with respect to their relation to the target. If the cells derived from WBS patients are a model of WBS (or of some features of WBS), and if the cells derived from 7Dup patients are a model of 7Dup, how are we to account of the inter-relation between the two cellular models? The issue becomes only more important as we proceed.

Samples from 12 patients and 4 controls were reprogrammed using non-integrative methods[20]. Because nothing was known about the diseases in any early developmental

[19] Working on this project are Sina Atashpaz, Antonio Adamo, Giuseppe D'Agostino, and Matteo Zanella.

[20] Conventional reprogramming delivers the reprogramming factors through viruses which are not only inefficient, but must be silenced and cause insertional mutagenesis upon integrating the genome. Instead, we delivered the factors directly as mRNA molecules (as well as some enhancing micro-RNAs), which are

stage, we decided to study the transcriptomes of the patient and control-derived iPSC using mRNA-sequencing. Among the first observations we could make, by far the most striking was that already at the iPSC stage we noticed differential expression, between patient- and control-derived cells, of brain-related genes. However, when comparing each disease to the control independently, these genes were few in a large and very heterogeneous set of a thousand of several hundreds of differentially expressed genes. Instead, when restricting our gaze to the *intersection* of genes independently found differentially expressed in *both* diseases, they showed striking enrichment in Gene Ontology categories that were *all* directly related to major elements of the disease phenotypes. Enriched categories were in three major branches of the Gene Ontology hierarchy of biological processes, recapitulating the major aspects of the disease phenotype: categories were either related to neuron differentiation and central nervous system development, to cardiovascular development (WBS patients often suffer from cardiovascular problems such as arterial sclerosis), or to migration (which should at least be involved in the cranio-facial features).

There are two observations that I wish to make here. First, despite the fact that these patients seem to undergo a relatively normal embryogenesis, one can see hints of the disease phenotype already at the pluripotent stage. This corroborates the point made earlier about subclinical phenotypes and late-onset diseases (section 5.3.3). Second, it also shows the importance of modeling the two diseases together: when looking at the genes deregulated in a single disease, the disease-relevant biological processes are flooded in a wide variety of Gene Ontology categories, and become visible by contrasting and intersecting the two models. This is yet another example of the point made in section 5.3.1 that differences between models can be harnessed to distinguish relevant features from background variation.

In order to follow these disease-related differences throughout development, we have differentiated these iPSC cells to two lineages that we expect to be particularly affected by the disease: neural crest stem cells, which are expected to be involved in the crano-facial phenotype, and cortical glutamatergic neurons (via neural progenitors), expected to be involved in the cognitive-behavioral phenotype. The rationale for studying such cell types is simple, and emphasized in a recent review:

transient, integration-free, and have a number of practical advantages including a better efficiency.

> "Because molecular pathways are shaped by cell-type-specific gene expression, it is preferable to study the molecular basis of a particular disease in the affected cell type." (Merkle and Eggan 2013)

While there is no denying this point, I wish to highlight that the 'affected' cell types are not the only relevant ones for the study of the diseases. Precisely because molecular pathways differ between cell-types, it is also important to study allegedly 'unaffected' cell types, such as iPSC. For even the absence of phenotype in non-affected cell-types itself calls for an explanation and can point to compensation mechanisms which not only yield insight in the disease-relevant mechanisms, but hint at possible therapeutic interventions. Such knowledge cannot come from the study of a single cell-type, but must come from their comparison. In this context, the different cell types derived from patients and controls are not simply modeling different aspects of the disease, but are providing contrastive spaces of representation in which disease-relevant interactions can stand out. Once more, the different cell types are used *together* for the study of the diseases in a way that neither of them could afford on its own.

5.4 Distributed modeling

5.4.1 Against the dyadic view

In the remainder of this chapter, I would like to use the iPSC research paradigm presented to make some remarks which extend to biomedical modeling more generally. To this end, it is useful to start by revisiting the diagnosis described in Chapter 2 as the 'dyadic and uni-directional view of modeling' (section 2.4.4). This view rested on three major assumptions which, I argued, have hampered much of the discussion around biomedical models (section 2.4. Taking them one by one in the context of iPSC modeling will suggest an alternative view of modeling, of which I would like to draw some implications.

Surrogacy

That model-based research replaces to some extent human experimentation does not imply that the model is "a surrogate for a human being" (ILAR and NRC 1998, p.10). In its simplest and strongest form, to view biomedical models as surrogates (section 2.4.1)

is to pretend that it is a human. I have shown in Chapter 3 that biomedical models are sometimes used in ways that have no meaningful counterpart in the target system, but even when models are used as replica (section 3.4.2), they are not simply replacing human beings. I argued in Chapter 1 (section 1.3) that even when scientists are testing in mice something as simple as the carcinogenicity of a substance, it would be inaccurate to say that the experimental setting is the same except for the replacement of the human with a mouse: the dose, as well as the endpoint, must also be translated between settings. Likewise, one does not, for instance, look for memory losses in an *in vitro* model of Alzheimer's Disease. Inevitably, the model presents something different from the target system, and this discrepancy must not only be acknowledged, but harnessed. For the value of *in vitro* models – and of most biomedical models – is not in replicating, but in reproducing in a different form: in offering contrastive spaces of representation.

The model-target dyad

Figure 5.4 above represents at least three different systems: the patient, the dish, and the animal model. A consequence of this is that we cannot understand modeling as a dyadic relationship between a model and a target (section 2.4.3). Indeed, the dish and the animal models are two independent models rather than the same model, because they can autonomously map to corresponding elements in the patient (the criterion proposed in section 1.3.1). Nevertheless, they are used in a common and integrated modeling effort: their findings do not simply add up, but are synergistic because they mutually reinforce each other. This distinguishes this kind of modeling from the mere juxtaposition of models with different scopes. Consider, what Weisberg, for instance, calls 'Multiple-models idealization': "the practice of building multiple related but incompatible models, each of which makes distinct claims about the nature and causal structure giving rise to a phenomenon." (Weisberg 2013, p.103) Weisberg explicitly refers to Richard Levins' famous paper on model building in population biology (Levins 1966). On Levins' account, models necessarily involve a trade-off between different desiderata (for instance generality and precision)[21], and as a consequence he argues that (in population biology) one should rely on

[21] As Orzack and Sober (1994) point out, in fact there is not necessarily such a trade-off, but we must grant to Levins that there is very often, if not most of the time, such trade-off.

complementary models. Distributed modeling, however, is not the same as having multiple models from which to choose according to particular situations or desiderata. Instead, the different model systems work *together in a way that each model could not afford on its own*.

This kind of structure is not new to iPSC disease-modeling, but instead represents a recurrent pattern in biomedical research. The same distributed nature of the modeling relationship is for instance visible in the establishment of cell types (see figure 5.2), and I have argued in section 5.2.3 that the validity of each *in vitro* cell type depended on both its phenotypic correspondence with the *in vivo*, and on its corresponding position in the sequence of models. I therefore suggest that we see modeling not as a relation between two systems, but as a network structure encompassing multiple systems, *including theoretical ones* (Chapter 4). This is what I propose to call distributed modeling. The use and meaning of any model system depends on its position in this network. Recall that this point was already made in my theoretical discussion of representation (section 1.1.2): whether a term successfully represents *as* depends on its position and relationships in a larger system of representation.

Unidirectionality

The very idea of replacing a dyadic view of modeling with a modeling distributed across a network of systems forces us to abandon thinking of modeling as a directional process (section 2.4.1). As figure 5.4 above shows, extrapolation – or better said, translation – runs bidirectionally between each node of the network, including the patient or clinical system. Indeed, the laboratory phenotypes purified are themselves 'extrapolations' of clinical observations (section 5.3.2; see also section 2.4.2). This means that the clinical should not be perceived as the mere endpoint of a research process, and therefore that we should abandon a unidirectional view of modeling. This changes the relationship between the lab and the clinic and dissolves the border between the two: the patients are but one system among others in this network of models.

This should not come as a surprise: in Chapter 1, I argued that there is no clear sense of modeling that would apply to biomedical models but not to the context of clinical research.

Participants in clinical trials are generally idealized (mostly due to exclusion rules), and the interventions made on them are modulated to epistemic purposes (see section 1.2.3). Clinical research is also modeling, and this is all the more obvious in contexts in which the patients represent something which they do not clearly instantiate (section 5.3.1). The clinical is part of the back and forth transactions between models, in other words between representational systems.

The notion of translational research should therefore be revisited. One cannot say, as the editor of the Journal of Translational Medicine did, that "the purpose of translational research is to test, in humans, novel therapeutic strategies developed through experimentation" (Marincola 2003, p.1). Instead, translational research is about creating bridges between the clinical setting and other models, in both directions and with the Bedside-to-Bench path not limited only to *ex vivo* materials and problem agenda. It means for lab scientists to translate clinical endpoints, biomarkers, and assays, to their biomedical models, for all scientists – including those involved in clinical research – to do their job with interoperability in mind.

5.4.2 The role of the patient

To consider patients as one model system among others implies an important change in the status of patients, which is most visible in the context of personalized medicine.

Part of the hype surrounding iPSC disease modeling comes from its natural suitability to personalized medicine. However, the whole discourse on personalized medicine is intrinsically ambiguous, oscillating between a quasi-nominalist interpretation – a medicine for each patient – and simply a narrower patient stratification. The same ambiguity can be found in the iPSC field, and is for instance present in many representations of the iPSC research pipeline, such as Figure 5.3 presented earlier. In the figure, an arrow goes from the 'New drugs' to the 'Patient', but who is this patient? Is it a single patient, or the representation of a group of patients? And perhaps more importantly, is it the same patient at the beginning and at the end of the process?

Consider the following (optimistic) passage from prominent stem cell biologists:

> "imagine, in the long run, that it will become routine not only to access the

> complete genetic information of a patient but to directly probe the patient's own iPSC-derived tissues for a broad range of medical questions." (Lee and Studer 2010, p.27)

Taken in isolation this extreme idea of personalized drug screenings performed on patient-specific iPSCs might seem unlikely for practical reasons, especially in the context of publicly funded healthcare[22]. In order to be cost-effective, this approach has to be coupled with another dimension. Consider, for instance, the following scenario:

> "iPS cells can be generated from any human who is taking a medicine. Thus, any effect or lack of effect of a particular drug that is detected during clinical treatment can be re-analysed using iPS cells from patients." (Nishikawa et al. 2008, p.727)

The effect (or lack of effect) is re-analyzed in iPSC cells, in order to understand it – but not for the patient herself, who has already experienced whether the drug will work for her or not. In this case, the patient at the beginning and at the end of the pipeline differ.

The project of personalized medicine is to bring both elements together in an iterative biomedical platform. For example, imagine that iPSC-based models have allowed us to identify *in vitro* predictors of drug response. For each patients, iPSC are derived to see whether they are likely to respond or not. Whether they in fact do, or not, feeds back into the predictors to improve them. In this context, every patient is at the same time a *source* and a *target* of extrapolation[23].

Critiques of iPSC modeling have picked on scientists' usage of medically loaded concepts in the context of the dish, starting with the very expression 'diseases-in-a-dish'. Some critics have complained that the research program would move the normative aspect of medicine to the lab: "The problem of disease is re-imagined as a technical matter to be understood (and resolved) in the laboratory." (Saha and Hurlbut 2011, p.E3) Although genuine, this old worry applies to all of biomedicine, and focusing on it obscures the opposite phenomenon: that of bringing the lab – in its epistemic rather than technical form – into the clinic.

Humans have always been both sources and targets of extrapolation. However, for most patients these two moments are separated, and participants in a clinical trials are

[22] In a private context, this is another story. In the US one can already pay a company 15,000 USD to have cancer drugs tested on xenografts of his/her own tumour – see for instance `www.championsoncology.com`.

[23] Note that the same logic applies to direct-to-consumer genetic testing: although consumers pay for the knowledge (or candidate knowledge) they receive about themselves, the business plan requires that they in turn serve as a basis for knowledge acquisition.

well aware of their change in status[24]. In the context of personalized medicine, the same patient is (imagined to be) both at the same time. Proponents of so-called 'Precision medicine' push this even further. Noting that the translational 'gulf' between the bench and the bedside is a structural problem intrinsic to our model for biomedical research[25], these authors are proposing to "conduct such research at 'point-of-care' in conjunction with the routine delivery of medical services." (Committee on a Framework for Development a New Taxonomy of Disease 2011, p.4) Here, the patient as a receiver of care is fully conflated with the patient as research participant. I do not wish to say that this is a problem, for this is also an opportunity for the patient who will benefit, if not from a fully personalized medicine, at least from a medicine tailored for a very precise patient group. It is, however, a dismissal of the separation between research and treatment that has been fundamental to medicine of the second half of the 20th century.

5.4.3 Structures of models as a locus of evaluation

To view modeling not as a relation between two systems, but as distributed among a network of systems, suggests that an evaluation of biomedical models should instead be an evaluation of whole modeling networks. However, if these configurations change from case to case, their evaluation might not be of much use. A good strategy would therefore be to strike a balance between the two extremes, and use as objects of evaluation modeling structures which are relatively stable across usages: not necessarily whole networks, but parts thereof or patterns which are recurrent enough for the evaluation to be useful. Here, I will only sketch an example of what such a structure could be.

Typically, human experimentation is not permissible unless we have good pre-clinical evidence, but clinical observations (from uncontrolled studies) can be gathered much more easily. An obvious problem with observational data is that it can only be correlative: controlled intervention is required to establish causality. However, if the same correlation is

[24] This point might however be questioned. Claude Bernard considered all medical interventions as human experiments, and as Canguilhem pointed out, it is ultimately only the medical practitioner who knows when his intent is therapeutic, and when it is epistemic – "Lui, et lui seul, sait précisément à quel moment l'intention et le sens de son intervention changent." (Canguilhem 1965, p.36)

[25] "Instead of moving clinical data and patient samples to research groups to allow analysis, the molecular data of patients should instead be directly available to researchers and health-care providers." (Committee on a Framework for Development a New Taxonomy of Disease 2011, p.61)

established for instance in an animal model, causality can be tested there, thereby providing a very strong indication of causality in the human system. The most common example of this practice is the testing, in animal, of the causal import of genetic variations found to be associated with diseases or phenotypes through genetic epidemiology such as GWAS. Another interesting example of this general strategy is the 'murine co-clinical trial' reported in Chen et al. (2012). On top of testing the overall efficacy of a treatment, the clinical trial also aimed at identifying biomarkers predicting differential treatment outcome. When the first candidates came out, bench scientists designed a parallel 'co-trial' in genetically engineered mice harbouring some of the markers for differential response, thus testing by controlled intervention whether the marker really makes a difference to the response, or whether the correlation was spurious (it did make a difference). It is interesting, in light of the previous point on the unidirectionality of the dyadic view, that in this example it is a prediction coming from the clinical setting that was ultimately put to test in the laboratory, rather than the opposite.

In a similar way, it is often the case that another organism can support a dynamical reconstruction of static knowledge about the organism of interest. It was for instance possible in model organisms to characterize the development of the central nervous system by tracking, in a combination of *in vitro* and *in vivo* experiments, differentiation processes. While *in vitro* experiments are equally available in humans, *in vivo* experiments are typically limited to post-mortem studies, offering only static pictures of the cell populations of the brain. However, if a given differentiation process was shown in the mouse, and the *in vitro* projection of this process is preserved in human *in vitro* experiments, chances are that the same process in happening in *in vivo* humans. Note the central role of the cell type in this structure: it is it which allows the bridge between species. An adequate account should situate notions such as these in the distributed modeling network.

It is typical of *in vitro* systems to work as translational systems, most often between model systems, and the modeling structure presented here is but one example of an efficient modeling structure which can be instantiated in virtually any context. It is for this reason that such structures are worth being evaluated.

5.4.4 Connectivity

That structures of models, rather than single model systems, can be the object of evaluation does not exclude single model systems from being evaluated on their own. However, the distributed account of modeling does suggest that they ought not be evaluated simply on their relationship to humans or human diseases (or the target system more generally), but also (and perhaps more importantly) on what they can bring to a network of distributed modeling. The importance of instrumental models (Chapter 3) emphasizes, for instance, reproducibility and internal validity over an external validity which is largely modulated by the model's position in the network. But even for other biomedical models, their relationship to other model systems has to be taken into consideration. First because some models are particularly fruitful in conjunction, but also because some models might have a greater capacity than others to form networks – what one might call connectivity.

A similar point was already made regarding model organisms. The NRC's 1985 report on biomedical models discusses what they call 'high connectivity models': "Such organisms may be regarded as high connectivity models, i.e., there is a great potential that connections will be made between observations on the model system and data on other systems." (Committee on Models for Biomedical Research 1985, p.51) Interpreted in light of their framework centered around a 'matrix of biological knowledge' (mentioned in section 4.2.3), this connectivity amounts to the sheer mass of knowledge about a given organism: the more is known about an organism, the easier it is to extrapolate some of this knowledge to other species such as human. The reason is most clearly visible in Daniel Steel's (2008) analytical account of extrapolation.

Steel proposes an interesting flavor of reductionism which he calls 'corrective asymmetry'. It is reductionist in the sense that it rests on "the idea that some levels of explanation are more fundamental than others" (Steel 2008, p.137), yet it implies neither logical reductionism, nor sufficiency in any way. Instead, "corrective asymmetry means that resources from the fundamental level are necessary to correct explanations provided at other levels, *but not vice versa*." (Steel 2008, p.137, original emphasis) This does not mean that the more fundamental level *always corrects* explanations or generalizations from the other levels, but that it sometime does, and that corrections do not run the other way around. Steel

argues that the way mechanistic knowledge supports extrapolation rests upon the plausible "assumption that mechanisms are correctively asymmetric with regard to the claims of interest to the extrapolation." (Steel 2008, p.126). When this is the case, mechanisms can explain (and predict) the breakdown, in some situations, of the generalization they underpin. Steel's proposal therefore begins with what he calls 'comparative process tracing':

> "The above discussion suggests a procedure for extrapolating a mechanism found in the base population to the target population, a procedure that I call *comparative process tracing*. First, learn the mechanism in the model organism, by means of process tracing or other experimental means. [...] Second, compare stages of the mechanism in the model organism with that of the target organism in which the two are most likely to differ significantly. [...] In general, the greater the similarity of configuration and behaviour of entities involved in the mechanism at these key stages, the stronger the basis for extrapolation." (Steel 2008, p.89, original emphasis)

The obvious problem with this strategy is that the comparison requires so much knowledge about the target system that, would this knowledge be available, it would screen-off the extrapolation. This is what Steel calls the 'extrapolator's circle': "establishing the suitability of the model would require already possessing detailed knowledge of the causal relationship in the target, in which case extrapolation would be unnecessary." (Steel 2008, p.4) The challenge, then, is to show how extrapolation can be supported in this way without comparing all the causally relevant elements of the model and target systems. Steel suggests "two key strategies for reducing the nodes of stages of the mechanisms that need to be compared in the model and target" : use background knowledge to focus "only on the points of likely relevant difference", or alternatively focus "on mechanism activities and components that are downstream in the sense of being more direct causes of the outcome" (Steel 2010, p.1060). While he does not give clear guidelines as to how the points of likely difference are to be identified, there are a variety of sources from which this can be derived. Mechanistic topology – the presence, for instance, of hubs or rate-limiting factors in a pathway or network – can suggest important way-points, and when their variation is correlated with the endpoint they can identify the key nodes in a mechanism.

Steel's strategies are certainly at play in biomedical research, and they explain part of the snowball effect of model organisms: the more knowledge we have about an organism, the easier it is to identify the conditions in which extrapolation from it are likely to break

down (the accumulation of technological resources and other infrastructures tied to the organism explains the rest, and arguably most of the snowball effect). Steel's account, however, falls into the dyadic view of modeling, for it considers modeling and extrapolation as a (directional) relationship between two isolated systems. While the NRC's notion of a matrix does acknowledge many-to-many modeling, it is strictly conceived of as a relation between corresponding *facts* in different species. Another way to interpret the snowball effect of model organisms was suggested in my discussion of the example of the zebrafish 'biosensor' (section 3.2.3), where I have argued that large-scale repositories of phenotypes can act as a map on which it will be possible to locate phenomena observed in future experiments. The recognition of these spaces of representation imply that Steel's account explains only a part of why (and how) knowledge of the model improves our capacity to extrapolate from it.

Finally, I would like to suggest is that there are also *experimental* kinds of connectivity. These sometimes come from apparently trivial technicalities: for instance the stocks of available antibodies for zebrafish are very poor, which is not only an intrinsic limitation of the system, but also makes it difficult to connect it to protein-level studies in other systems. Sometimes, these experimental connections are central to a research programme, and transplantation, which repeatedly came back in the present work, is one such example.

5.5 The new model organism?

As pointed out in the introduction (section 0.3.1), the development of tissue culture prompted a reconsideration of the ethical issues surrounding animal experimentation. Even Rheinberger recently wrote that

> "we are entering in an epoch of the eclipse of animal models in medicine and human research. The mode of reaching the molecular level in the test tube and the subsequent molecular re-cellularization of research has created the perspective of an investigation of human specificity without intermediate animal models." (Rheinberger 2011a, p.168)

Confronted with this possibility, most scientists have argued that the *in vitro* and the *in vivo* "are most often complementary rather than substitutes for each other" (Institute of Laboratory Animal Resources (ILAR) 1977, p.67). That these models cannot simply be

substituted for each other follows from the fact that they are seldom mere surrogates for patients. If there are cases of *in vitro* systems replacing animals, it will most likely be as measuring devices (Chapter 3) for some very specific purposes. This is for instance the case in the recent development of the *in vitro* ScoreCards to replace teratoma assays for the assessment of pluripotency (Bock et al. 2011). A lot of animal experimentation could thus be replaced, although a lot of work would be needed for this to be implemented – work for which animals will be of fundamental importance. To this end, it will be necessary to develop the right incentives, not so much through restrictions of animal usages but through targeted funding of alternatives.

This being said, I have tried to show throughout this chapter that the *in vitro* and *in vivo* are often not only complementary (which would be expected from the simple fact that they differ in the way they model), but that they are also synergistic – that they get much of their value from being used together. For the very meaning of the expression *in vitro*, as well as its productivity, derives from its incompleteness relative to the *in vivo*, or more precisely from its ability to make contrasts that are generally impossible *in vivo* (section 5.2.4).

The *in vitro* cannot replace the *in vivo* entirely, because its meaning and existence are tied to it. However, a promise of iPSC-based research, and in fact of much of personalized medicine, is precisely that patients themselves will provide the *in vivo* counterpart. In a 2008 meeting of the American Association for Cancer Research, Sydney Brenner claimed that "[w]e don't have to look for model organisms anymore because we are the model organism" (quoted in Ledford 2008, p.843; Goldstein 2012 makes a similar claim in the context of iPSC disease modeling). This is a particularly strong statement coming from someone who won the Nobel prize for work on *C. Elegans*. Humans do have major characteristics of model organisms: accumulated knowledge and technologies, research communities and infrastructures, and even an emphasis on organismal understanding (Ankeny and Leonelli 2011); where they depart from model organisms is that they are not generally taken to be representative of other species. Whether or not we ought to call them model organism, Brenner's message is that the clinical setting offers so much information that, in the right conjunction with the lab (especially with the *in vitro*), is sufficient for biomedical research

(or at least, given the context from which he was quoted, sufficient for cancer research). Indeed, both iPSC and biopsies have the important characteristic of coming with the clinical history of the patient, thus providing an effective *in vivo-in vitro* couple without apparent need for other organisms. Furthermore, the development of relatively non-invasive tracing technologies have opened up a whole new range of clinical observations.

While this is certainly an important transformation occurring in biomedical research, it is clearly too early to toll the bell of experimental organisms. For one thing, the sheer amount of tools and techniques available for animal experimentation remains unparalleled. More fundamentally, a couple of more fundamental limitations must be kept in mind. First, as mentioned earlier (section 5.4.3), while the range of clinical observations may dramatically increase, it is in general not possible to make them in the context of controlled interventions, and as a consequence they do not allow one to establish causality. Of course, clinical trials represent an exception to this, and in this respect, cancer is a special case, for many trials are not about testing new treatments, but allocating already accepted treatments to selected subpopulations of patients. In other contexts, substantial pre-clinical evidence is required in order for a clinical trial to be permissible. In the absence of animal studies, this evidence can come only from correlative clinical observations, and *in vitro* models which are at risk of producing artefacts. To some extent, the two sources can correct each other's flaws, for if a causal relationship is an *in vitro* artefact, it is unlikely to have a corresponding *in vivo* correlation. Furthermore, many problems can be decomposed in such a way that is amenable to complementary *in vivo* and *in vitro* studies. For instance, a new strategy in toxicology is to give patients prospective drugs in extremely small amount, in order to assess basic pharmacokinetics, for instance to see where the drug is going in the body. Toxicity can then be estimated *in vitro* on the tissues that were identified in this first step.

This is, however, already quite far from the therapeutic intent that are supposed to underpin the practice of clinical trials. Some scientists have indeed voiced their concern over what they regard as a recent trend to replace pre-clinical with clinical studies, such as the debated phase-0 clinical trials. This is reminiscent of the warning issued more than a century ago by the National Academy of Science, writing that if animal experimentation is restricted, "medical science will continue to advance [...] but there will be this important

difference, that the experimenters will be medical practitioners and the victims human beings." (Committee on the Use of Laboratory Animals 1988, p.91). Ultimately, the question depends on the amount of risk which we, as prospective patients, are ready to take.

Conclusion

In Chapter 1, I have argued that there is no robust way to distinguish biomedical modeling from direct experimentation. Extrapolations made from biomedical models to human or clinical systems are not different in nature from those made between different experimental systems. Indeed, I have argued along Bernard that biomedical modeling and experimental medicine both have the same conditions of possibility (section 4.2.1). For this reason, I have proposed an epistemological analysis of biomedical modeling which is at the same time a philosophy of experimentation, and which does not presuppose a distinction between the two[26].

Both biomedical models and experimental systems more generally rely on representation, and another important claim made in Chapter 1 was that representation is not merely a relation between the representing and the represented, but that this relation is relative to a broader representational system. A simple way to illustrate this point is through the metaphor of *translation*, which is a much more appropriate notion than that of extrapolation. Extrapolation originates from mathematics (in opposition to interpolation), where it means to infer, on the basis of the data points we have, a point outside the range of our dataset, typically on the basis of a functional relationship. Underlying the idea of extrapolation is therefore a sort of continuity, and although this idea fits very well the logic of simple toxicology scaling rule, I believe it is generally ill-suited for the context of modeling. *Biomedical models are not simply scaled-down versions of their target, but they are the projection of their target in a different space of representation.* And this is exactly what the notion of translation is supposed to capture. Translating a word does not mean to find

[26] As noted in the introduction (section 0.1), I am concerned here with the use of material, biological models such as those that have by and large been the topic of the present work. Mathematical or abstract models, as well as Weber's 'experimental modeling' (Weber 2012), would most likely benefit from a distinct analysis.

the equivalent of this word in the target language – words seldom (if ever) have such an equivalent. Instead, the translator is forced to ask for the context in which the word is used: the translator is translating not a single word, but something broader and which includes the relation of the word with its context. So is, I have argued, the modeler.

This view of biomedical research has a number of implications which I have tried to highlight, and which contradict commonly-held assumptions about modeling (Chapter 2). The first assumption – that biomedical models as surrogates for human patients – is perhaps the most obviously problematic. As I argued in Chapter 1, a biomedical model is not limited to an organism (or to cells in a dish), but generally contains a number of other components which are modeling as much as the organism is (section 1.3). In Porsolt's forced-swim test, the cylinder full of water stands for the depressing environment, LPS stands for an infection, etc. (section 1.3). As a consequence, if there is surrogacy, the model at least encompasses more than the animal (or cells) and stands for a larger clinical setting, rather than just for the patient. Conducting biomedical research on model systems does not therefore imply the assumption that these "will respond to manipulations in the same way as the target would, if it were examined directly." (Bolker 2009, p.490) Indeed, even when animals have appeared to be used as surrogates, the meaning of 'responding in the same way' (as well as 'to the same manipulations') has to be qualified in important ways[27].

As I have shown in Chapter 3, some other cases do not fit at all the surrogate role of biomedical models, and forcing them into such a conception of modeling misrepresents their use and the function against which they ought to be evaluated. Instead, they are used as instruments, and I have identified different senses of this instrumental role (section 3.3.1), concentrating on biomedical models used as measuring devices. Measuring devices locate their target on a constrained space of representation. It is constrained in several ways: a space has a structure with defined (and limited) dimensionality, and very often a digitalization (section 3.3.2). The recognition of this role first shows that although biomedical models ultimately aim at informing us about human biology and pathology, they do so in a variety of ways, and it is with respect to these proximate functions that they must be evaluated.

[27] Some important examples to this effect were presented in Chapter 2, such as the early endpoints in the CCNSC (section 2.3.2), and Willner's example of stimulant-induced stereotyped behavior (section 2.4.1).

The instrumental role suggests a different way of looking at biomedical models, according to which each instrument or model system is tied to a representational space. In this view, there is not a simple one-to-one mapping between the model and the target, but a variety of representational spaces related through theoretical relationships. For this reason, I have explored in Chapter 4 the interplay between biomedical models and theories. I have argued that this point can inform a contemporary debate in cancer research (section 4.3.4) by replacing an obsession with phenocopy with an attention to how an instrument or model locates its target within a theoretical framework. In this context, a key task of epistemological analysis should be to understand how given spaces of representation interlock, or fail to do so.

Situating this view of biomedical modeling within a coherentist view of science (section 4.4) suggests that it is not best seen as a transfer from a model to a target, but rather as a "shuttling back and forth between different spaces of representation" (Rheinberger 1997, p.108). This is what I have tried to capture, in Chapter 5, with the expression of distributed modeling. According to the criterion proposed in Chapter 1 (section 1.3.1), two elements are part of the same model insofar as their modeling capacity is interdependent. In Chapter 5, I have shown that what are generally conceived of as distinct biomedical models are very often parts of a distributed model which, according to this criterion, ought to be considered together as a single model, for their modeling capacity is strongly interdependent. This is most obvious in cases involving *in vitro* models. The point is once more suggested by the translation metaphor, in which the meaning of a word depends of the other words in the sentence. Nevertheless, there might still be a value in individuating what are held to be distinct biomedical models for the purpose of epistemological analysis, especially since they represent relatively stable modules that can be reassembled into new experimental configurations and thereby represent useful loci of evaluation. However, in doing so one must be careful to pay attention to the transfers between these sub-models. Indeed, perhaps the greatest flaw of the dyadic view of modeling is that it ignores the relationships between different models (see section 2.4.3).

I have highlighted two important consequences of this distributed picture of modeling for the purpose of evaluating biomedical models. I have argued that larger modeling structures

– robust configurations of sub-models – should also be considered as loci of evaluation (section 5.4.3). I have also suggested that some models (systems generally conceived as distinct models) might have a greater connectivity than others, in the sense of a greater capacity to mesh with other models and to form networks of distributed modeling (section 5.4.4).

We must also cease to consider the target system as something outside of this network, to which facts are extrapolated once they have been worked out in the model (section 2.4.2). I have instead argued in Chapter 5 that the patient is his/herself a sub-model in the distributed network of modeling, and that transfers from and to the patient are not different in nature from those happening between other biomedical models. This has consequences for the status of the patient which ought not be ignored (section 5.4.2), but it also bears on the evaluation of biomedical models, for which it is necessary to consider a model's ability to include, and respond to, inputs from the clinical system.

To consider humans as a model is the final step in the deconstruction of a rigid distinction between modeling and experimentation, and in the realization that the problems of modeling are the problems of experimental medicine. However, it would seem (and rightly so) that despite the changing world of biomedicine, humans still have a special role with respect to biomedical models: it is after all, their interests that are expected to be served by biomedical research. In what remains, I would like to address this issue.

6.6 Modeling in applied research

Throughout this work, and especially in Chapter 4 (section 4.4), I have taken a particular stance regarding models and their target, which can be summarized in the following passage from Rheinberger:

> "Neither models nor reals are givens; models do not stand for absolute referents. Models are simply the privileged objects of manipulation. They become privileged, not by the things themselves they pretend to model, but through comparison with other model systems." (Rheinberger 1997, p.91)

I believe the basic Kantian point according to which we always apprehend reality as a representation is uncontroversial. In Chapter 4, I have argued that this constructive aspect gives the scientist some freedom in designing representational systems that are useful in our dealings with the world. This translates into a flexibility of biomedical models and the way they connect with other systems, for all that is required of a biomedical model is that it provides a projection of the target in a space of representation that relates in useful ways with other representations. In general, the biomedical model itself is not the only space of representation with which we can tinker: just as the model can be brought into alignment with the rest of our representations, so can the rest of our research systems be modified to accommodate connections to and from the model.

This being said, one might object (and rightly so) that even in the absence of absolute referents, not all our representations are equally flexible. Some are more entrenched than others, or more tightly connected with the rest of our knowledge, but constraints can also come from non-epistemic factors. To different extents in different contexts, society sets science an agenda which, the thought goes, provides some fixed points in the representational systems of science[28]. This is especially obvious in sciences such as biomedicine, which explicitly aims at improving our management of health and disease. In contrast, it does seem that some more fundamental areas of the biological sciences have aims that are more vaguely defined, hence are free to use the tools that are the most productive and let these tools determine the directions in which science goes. Obviously, the relation always runs both ways – experimental systems always influence research questions, and research questions always to some extent constrain experimental systems – but sometimes one side has more weight than the other. For epistemological analysis, this difference in degrees of freedom is perhaps a more interesting way to render the traditional distinction between applied and 'pure' science.

[28] This point is raised by John L. Fuller and William R. Thompson (1960) in their foundational book on the genetics of behavior: "The distinction between human and animal behavior genetics is more than a matter of species studied or the techniques which are feasible in the two fields. The primary objectives of the workers in the two areas are different. Animal experimenters use genetics as a device to study the nature of variables which determine behavior. In gathering such information, traits and subjects are selected for study because of experimental convenience [...] In contrast, workers in human behavior genetics have concentrated on problems of social significance [...]" (Fuller and Thompson 1960, quoted in Wahlsten 2012, p.476) Society, these authors suggest, is interested in some questions (such as the sources of variations in intelligence) independently of whether these phenomena (or the very notion of intelligence) are experimentally tractable.

Nevertheless, I wish to emphasize that representations we perceive as 'given' by society are much more flexible than we tend to think. Even clinical endpoints, in a context as concrete as oncology, can be represented in very different ways. The very notion of 'cure', or 'complete remission', is telling: a patient is generally considered cured if five years have passed without sign of cancer. Such an endpoint might be useful, but it is by no means the only possible endpoint of cancer treatment, nor the only possible interpretation of a cure.

When Harold E. Varmus became the current director of the NCI in 2010, he held a speech sketching his vision and priorities for what was to come. He said that "[w]e need to think a little more clearly about how we frame the questions that we're trying to answer", which are "not as simple as 'How do I cure cancer?' ":

> "it's getting to a level of specificity that is *based on new developments in our science*, and look at questions that are not pie in the sky, but have a substantial prospect of answerability in the foreseeable future." (Varmus 2010)

Varmus consulted the scientific community to identify 'provocative, answerable questions' that applications to the NCI should address[29]. Although society entrusted the NCI with a mission, in the end it is the scientists who ask the questions they are to answer, and set the terms in which the answer will be formulated.

This is most obvious in the revision of disease or symptom categories, and an interesting example is the Research Domain Criteria (RDoC) of the National Institute of Mental Health, which are intended first and foremost for research purposes. More generally, the gradual transformation of nosology towards etiological and increasingly molecular categories is not some kind of convergence to 'true' categories, but first and foremost a way to make biological knowledge relevant to medicine. Herein lies the limitation of the translation metaphor in translational research. Translation suggests two pre-existing languages, and a pre-existing text to be translated. It suggests unidirectionality (section 2.4.2). Translational research, instead, is more like a dialogue, with the possibility of attuning both languages to each other.

[29] The questions are available at http://provocativequestions.nci.nih.gov/.

Appendix A

Test statistics and the coin-tossing argument

A.1 Fundamentals of test evaluation

There are several ways to assess the quality of a test or screening procedure, and some of them are more relevant than others depending on the broader goal of the procedure. In order to show this, it is useful to start with the basic terms presented in table A.1, which defines some of the terms used to qualify a given test result with respect to real states of the world (or, for practical purposes, to some gold standard) – in the present context whether a compound is effective, as attested by clinical trials.

	Compound is effective	Compound is uneffective
Test outcome is positive	True positive (TP)	False positive (FP)
Test outcome is negative	False negative (FN)	True negative (TN)

Table A.1: Terms qualifying the results of a screening procedure on the basis of their correspondence to 'real' or externally validated results.

True positives and true negatives are compounds which were successfully found by the test as, respectively, effective or not, whereas false positives and false negatives represent errors of the test – respectively type I and type II errors.

On the basis of these terms, one can define several other terms that are often used in statistical analysis of this kind:

- **Positive predictive value** (PPV): the proportion of true positives among all com-

pounds that the test gives as positive, i.e. TP/(TP+FP)

- **Negative predictive value** (NPV): the proportion of true negatives among all compounds that the test gives as negative, i.e. TN/(TN+FN)

- **Sensitivity** (also known as true positive rate): the proportion of true positives over all effective compounds, i.e. TP/(TP+FN).

- **Specificity** (also known as true negative rate): the proportion of true negatives over all uneffective compounds, i.e. TN/(TN+FP).

- **Accuracy**: the proportion of all compounds that are correctly recognized, i.e. (TP+TN)/(TP+FP+TN+FN).

A test with 100% accuracy necessarily has 100% in all four other values. However, high but incomplete accuracy can be misleading. Imagine, for example, a screening procedure which always gives a negative result. If the test is aimed at detecting a rare event obtaining in, say, 1/1000 of all cases tested, this will mean that the test is right 99.9% of the time (has a 99.9% accuracy). Nevertheless, for all practical purposes, such a test is useless.

Sensitivity and specificity are generally more useful. In the example given, the test would have a specificity of 100% but, because it never gives a positive, a sensitivity of 0% (0/1). The problem, however, is that they require information which we do not always have, namely knowledge of the proportion of compounds that are in reality effective or ineffective.

A.2 Base rates and the coin-tossing argument

The lack of proper knowledge of the the proportion of compounds that are effective/ineffective has the danger of leading to an assessment error more broadly known as the base rate fallacy. With some notable exceptions (Knight 2011), many critics of animal testing make little effort to prevent their readers from making this mistake. This is most obvious in the comparison with coin-tossing, which repeatedly shows up in the literature. According to critics, animal studies "predicted human response about as well as a coin toss" (Shanks et al. 2009, p.10; also p.6).

To see why this is to commit a form of base rate fallacy, let us imagine two worlds: in W1, 200 out of 1000 possible compounds are actually effective drugs for the condition of interest, while in W2, 2 out of 1000 compounds are actually effective. In addition, let us assume that animal experimentation has a PPV of 1/10, as suggested by the following passage quoted by Shanks et al.:

> "Currently, nine out of ten experimental drugs fail in clinical studies because we cannot accurately predict how they will behave in people based on laboratory and animal studies" (U.S. Secretary of Health and Human Services in 2007, quoted in Shanks et al. 2009, p.4)

We therefore have two possible drug discovery strategies:

- S1: A pre-clinical screening involving animal studies, where as earlier only 1 out of 10 compounds entering clinical study turns out to be positive.
- S2: A pre-clinical screening involving coin-tossing.

If we are in W1, we would toss coins for 1000 compounds and choose 500 for clinical study, of which on average 100 would actually be effective. In other words, 2 out of every 10 compounds entering clinical studies would be effective drugs, which would be twice as much than we get using animal studies.

However, if we are in W2, we would toss coins for 1000 compounds and choose 500 for clinical study, of which on average 1 would actually be an effective drug. That means that to find one successful drug, we would need to do 1 successful and 499 failed clinical trials. This is catastrophically worse than the 1/10 attrition rate after animal studies.

Some commentators have made this mistake even in contexts where much information was known. For instance, Shanks et al. (2009) discuss a case in which drugs of known human efficiency were tested back in animals. Nevertheless, they make the same mistake in assessing coin-tossing:

> "The animal tests were shown to have a sensitivity of 0.52 and the positive predictive value was 0.31. The sensitivity is about what one would expect from a coin toss and the PPV less. Not what is considered predictive in the scientific sense of the word." (Shanks et al. 2009, p.6)

The sensitivity of a coin toss should indeed be 0.5, for the coin toss will always give on average an equal number of True positives and False negatives, which leads to

$$Sensitivity = \frac{TP}{TP + FN} = \frac{x}{x + x} = 0.5$$

However, the same cannot be said of the positive predictive value, because it depends on the base rate. Depending on whether we are in W1 or W2, we will have the following:

$$PPV_{W1} = \frac{TP}{TP + FP} = \frac{100}{100 + 400} = 0.2$$

$$PPV_{W2} = \frac{TP}{TP + FP} = \frac{1}{1 + 499} = 0.002$$

It both cases, this is significantly lower than the reported PPV for animal studies. Although we do not know in which of W1 or W2 we actually live, we have reasonable grounds to assume that it is more like W2, and in fact most certainly much lower, for only a few rare compounds are actually effective drugs. And in this context, animal studies perform orders of magnitude better than coin tossing.

These problems do not arise in all comparisons with coin-tossing (see for instance Salsburg 1983), and there might be specific contexts where animal studies are systematically misleading. But one must be careful not to forget the base rate when interpreting seemingly bad attrition rates.

Appendix B

Biological and technical replicates

The expressions 'biological replicate' and 'technical replicate' are ubiquitous in biology, but they are surprisingly recent. At first glance, the adjectives suggest that one can understand these notions in terms of a distinction between biological variation and technological variations. This however becomes problematic as soon as biological systems are used as instruments. Consider the example of the teratoma assay, the gold standard for assessing the differentiation potential of alleged pluripotent stem cells. In this test, the cells are injected in a mouse to see if they will grow a teratoma reconstituting the three germ layers: if they do, then by extrapolation they should be able to differentiate into any cell type of the organism, and they are therefore considered pluripotent. Growing and differentiating are inherently biological processes, and yet if we performed two teratoma assays using cells from the same dish, we would generally call these experiments 'technical replicates'. The reason, I would suggest, is that the mice are here measuring devices just like those presented above, and are used to assess a property of the injected cells.

If one looks up the terms in the biomedical ontology of the National Institute of Health, the examples for biological and technical replicates are, respectively, "a patient in a given arm of a trial" and "aliquots of a tissue subjected to parallel assays". Likewise, in textbooks the distinction will hinge on whether the experiments are performed twice on the same sample, or on two different samples. 'Sameness', here, is accomplished by a sampling assumed to be random (see discussion in section 1.2.2).

Consider another example: a prospective stem cell therapy, in which we inject stem cells (coming from the same population) into the brain of two patients affected by a neurological

disease, in order to see whether the intervention alleviates the disease. Clearly, we would not consider the two experiments as technical replicates, despite the fact that both patients were injected random samples of the same population of cells. Instead, we would consider that it is the two patients that represent two samples, and therefore that the two experiments are biological replicates. But why, in the case of the teratoma assay, did we not consider the two mice as two samples? Or why, in the case of stem cell therapy, did we not consider the patients are measuring devices, meant to assess the regenerative properties of the injected cells?

An important difference is that in the teratoma assay, a robust link was already established between the pluripotency of the cells and their ability to grow teratoma under certain experimental conditions. As a consequence, the process of engraftment, growth and teratoma formation can be black-boxed, and the mouse can be used as a mere instrument for the investigation of the injected cells. In other words, the mouse and the assay can be considered as 'technical objects', while the injected cells are part of the 'epistemic thing' – that which we are currently trying to understand:

> "A technical product, as everybody expects, has to fulfill the purpose implemented in its construction. It is first and foremost an answering machine. In contrast, an epistemic object is first and foremost a question-generating machine." (Rheinberger 1997, p.32)

In contrast, in the case of the stem cell treatment, it is not (or not only) the stem cells that are the object of investigation, but the intervention performed with them. The intervention and the processes it involves cannot be black-boxed and used as a technical object, because they have not yet been established.

If this analysis is correct, then what in these circumstances are considered biological replicates could, in other circumstances, be considered technical replicates. Suppose, for instance, that a given stem cell therapy was shown to robustly improve a given pathological condition. However, in order to make the therapy more accessible and reduce risks of rejection, we are investigating whether induced pluripotent stem cells (iPSC) are as efficient as embryonic stem cells for this treatment: we compare the outcome of the treatment using different population of cells, reprogrammed through different means. In these circumstances, the treatment itself effectively becomes a measurement assay performed on

the different kinds of cells. Therefore, two patients injected with the very same population of cells should be considered as technical replicates. Biological replicates would instead be two patients respectively injected with two populations of cells which were reprogrammed using the same method. Because of this shifting way to perceive given kinds of variation, it is common for someone's technical replicate to be someone else's biological replicate. Therefore, instead of the distinction between biological and technical replicates, it might be more precise to speak, following Rheinberger's distinction, of epistemic and technical replicates.

Acknowledgments

I would like to begin this round of acknowledgments by expressing my most sincere gratitude to Giovanni Boniolo, for his (often under-appreciated) commitment and work in founding and sustaining the FOLSATEC program. I am also grateful to my colleagues and to the rest of the program's faculty. The program has offered me the most stimulating intellectual experience I have ever enjoyed. My sincere thanks also to the staff and scientists at the IFOM-IEO campus who have helped to make this possible, especially to the scientists that have adopted me in their group for a time and those that spent considerable efforts tutoring me: Luisa Lanfrancone, Marina Mione, Cristina Santoriello, Saverio Minucci, Iros Barozzi, and most importantly the members of Giuseppe Testa's lab.

My internal supervisor Giuseppe Testa has given me unique opportunities, especially by entrusting me with scientific responsibilities and investing in me, but also through his personal 'Fragestellung' which in many respects was new to me. I also very much benefited from the philosophical acuity of my external supervisor Marcel Weber, and from the discussions of the 'Lake Geneva Biological Interest Group' (lgBIG). I would also like to thank the participants and organizers of three workshops that were particularly enriching for me: the Second European Advanced Seminar in the Philosophy of the Life Sciences (EASPLS 2012) on models in biology; the workshop "Animal Models, Model Animals? Meanings and Practices in the Biomedical Sciences" (Centre for the History of Science, Technology and Medicine, University of Manchester, 2012); and the International Advanced Seminar in Philosophy of Medicine (IASPM, Paris 2013).

Very special thanks go to Fridolin Groß for his careful reading, and his most helpful comments. Finally, I would like to thank all those that have given me feedback on pieces of this work or discussed these issues with me, especially Thomas Reydon, Robert Meunier, Annette Kappeler, and Maël Lemoine.

Bibliography

Allegrucci, C., M. D. Rushton, J. E. Dixon, V. Sottile, M. Shah, R. Kumari, S. Watson, R. Alberio, and A. D. Johnson (2011). Epigenetic reprogramming of breast cancer cells with oocyte extracts. *Molecular Cancer 10*(1), 7.

Alley, M. C., M. G. Hollingshead, D. J. Dykes, and W. R. Waud (2004). Human Tumor Xenograft Models in NCI Drug Development. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 125–152. Humana Press.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement 10*, 67–78.

Ankeny, R. A. (2007). Wormy Logic : Model Organisms As Case-Based Reasoning. In E. Lunbeck, A. N. H. Creager, and M. N. Wise (Eds.), *Science without laws: model systems, cases, exemplary narratives*, Number 07, pp. 46–58. Duke University Press.

Ankeny, R. A. (2010). Using Cases to Establish Novel Diagnoses: Creating Generic Facts by Making Particular Facts Travel Together. In P. Howlett and M. S. Morgan (Eds.), *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, pp. 252–272. Cambridge University Press.

Ankeny, R. a. and S. Leonelli (2011). What's so special about model organisms? *Studies In History and Philosophy of Science Part A 42*(2), 313–323.

Armitage, P. and R. Doll (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer VIII*(I), 1–12.

Armitage, P. and R. Doll (1957). A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British Journal of Cancer 11*(2), 161–169.

Armitage, P. and M. Schneiderman (1958). Statistical problems in a mass screening program. *Annals of the New York Academy of Sciences 76*, 896–908.

Arrowsmith, J. (2012). A decade of change. *Nature reviews. Drug discovery 11*(1), 17–8.

Bailer-Jones, D. (2002). Models, metaphors, and analogies. In P. Machamer and M. Silberstein (Eds.), *The Blackwell Guide to the Philosophy of Science*, pp. 108–126. Blackwell Publishers Ltd.

Baird, D. (2003). Thing Knowledge: Outline of a materialist theory of knowledge. In H. Radder (Ed.), *The Philosophy of Scientific Experimentation*, pp. 39–67. University of Pittsburgh Press.

Bellin, M. and M. Marchetto (2012). Induced pluripotent stem cells: the new patient? *Nature Reviews Molecular Cell Biology 13*(11), 713–726.

Bellugi, U., L. Lichtenberger, W. Jones, Z. Lai, M. St. George, and M. S. George (2000). I . The Neurocognitive Profile of Williams Syndrome : A Complex Pattern of Strengths and Weaknesses. *Journal of Cognitive Neuroscience 12*(supplement 1), 7–29.

Belzung, C. and M. Lemoine (2011). Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. *Biology of Mood & Anxiety Disorders 1*(9), 1–14.

Bernard, C. (1865). *Introduction à l'étude de la médecine expérimentale*. Baillière.

Bhowmick, N. and E. Neilson (2004). Stromal fibroblasts in cancer initiation and progression. *Nature 432*(November), 332–337.

Biancotti, J.-C., K. Narwani, N. Buehler, B. Mandefro, T. Golan-Lev, O. Yanuka, A. Clark, D. Hill, N. Benvenisty, and N. Lavon (2010). Human embryonic stem cells as models for aneuploid chromosomal syndromes. *Stem cells 28*(9), 1530–40.

Blasimme, A. (2012). *Regenerative medicine and the governance of stem cell innovation*. Ph. D. thesis, European School of Molecular Medicine (SEMM) and University of Milan.

Blasimme, A., P. Maugeri, and P.-L. Germain (2013). What mechanisms can't do: Explanatory frameworks and the function of the p53 gene in molecular oncology. *Studies in History and Philosophy of Biological and Biomedical Sciences 44*(3), 374–384.

Blasimme, A., B. Schmietow, and G. Testa (2013). Reprogramming Potentiality: The Co-Production of Stem Cell Policy and Democracy. *The American Journal of Bioethics 13*(1), 30–32.

Bloche, M. G. (2004). Race-based therapeutics. *New England Journal of Medicine 351*(20), 2035–2037.

Bock, C., E. Kiskinis, G. Verstappen, and H. Gu (2011). Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell 144*(3), 439–452.

Bogen, J. and J. Woodward (1988). Saving the phenomena. *The Philosophical Review 97*(3), 303–352.

Boisclair, M. D., D. A. Egan, K. Huberman, and R. Infantino (2004). High-Throughput Screening in Industry. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 23–40. Humana Press.

Boland, M. J., J. L. Hazen, K. L. Nazor, A. R. Rodriguez, W. Gifford, G. Martin, S. Kupriyanov, and K. K. Baldwin (2009). Adult mice generated from induced pluripotent stem cells. *Nature 461*(7260), 91–4.

Bolker, J. a. (1995). Model systems in developmental biology. *BioEssays : news and reviews in molecular, cellular and developmental biology 17*(5), 451–5.

Bolker, J. a. (2009). Exemplary and surrogate models: two modes of representation in biology. *Perspectives in biology and medicine 52*(4), 485–99.

Bonnet, D. and J. Dick (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature medicine 3*(7), 730–738.

Boyd, M. R. (2004). The NCI Human Tumor Cell Line (60-Cell) Screen: Concept, Implementation, and Applications. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 41–62. Humana Press.

Boyd, M. R. and K. D. Paull (1995). Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen. *Drug Development Research 34*(2), 91–109.

Boyd, R. (1999). Homeostasis, species, and higher taxa. In R. A. Wilson (Ed.), *Species: New Interdisciplinary Essays*. MIT Press.

Brennand, K., A. Simone, N. Tran, and F. Gage (2012). Modeling psychiatric disorders at the cellular and network levels. *Molecular psychiatry*, 1–15.

Buchner, E. (1897). Alkoholische Gährung ohne Hefezellen. *Berichte der Deutschen Chemischen Gesellschaft 30*, 117–124.

Burian, R. M. (1993). How the choice of experimental organism matters: epistemological reflections on an aspect of biological practice. *Journal of the history of biology 26*(2), 351–67.

Burian, R. M. (1995). Comments on Rheinberger's "From Experimental Systems to Cultures of Experimentation". In G. Wolters and J. G. Lennox (Eds.), *Concepts, Theories, and Rationality in the Biological Sciences*, pp. 123–136. University of Pittsburgh Press.

Cambrosio, A. and P. Keating (1995). *Exquisite Specificity: The Monoclonal Antibody Revolution*. Oxford University Press.

Cambrosio, A. and P. Keating (2003). *Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine*. The MIT Press.

Canguilhem, G. (1965). *La connaissance de la vie* (2e ed.). Vrin.

Carette, J. E., J. Pruszak, M. Varadarajan, V. a. Blomen, S. Gokhale, F. D. Camargo, M. Wernig, R. Jaenisch, and T. R. Brummelkamp (2010). Generation of iPSCs from cultured human malignant cells. *Blood 115*(20), 4039–42.

Carrel, A. (1929). Physiological Time. *Science 74*, 618–621.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, USA.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Clarendon Press.

Cartwright, N. (1999). *The Dappled World : A Study of the Boundaries of Science*. Cambridge University Press.

Chamberlain, S. J., X.-J. Li, and M. Lalande (2008). Induced pluripotent stem (iPS) cells as in vitro models of human neurogenetic disorders. *Neurogenetics 9*(4), 227–35.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Chang, H. (2009). Operationalism. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

Chen, Z., K. Cheng, Z. Walton, Y. Wang, H. Ebi, T. Shimamura, Y. Liu, T. Tupper, J. Ouyang, J. Li, P. Gao, M. S. Woo, C. Xu, M. Yanagita, A. Altabef, S. Wang, C. Lee, Y. Nakada, C. G. Peña, Y. Sun, Y. Franchetti, C. Yao, A. Saur, M. D. Cameron, M. Nishino, D. N. Hayes, M. D. Wilkerson, P. J. Roberts, C. B. Lee, N. Bardeesy, M. Butaney, L. R. Chirieac, D. B. Costa, D. Jackman, N. E. Sharpless, D. H. Castrillon, G. D. Demetri, P. a. Jänne, P. P. Pandolfi, L. C. Cantley, A. L. Kung, J. a. Engelman, and K.-K. Wong (2012). A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response. *Nature*, 1–5.

Cherry, A. and G. Daley (2012). Reprogramming Cellular Identity for Regenerative Medicine. *Cell 148*(6), 1110–1122.

Churchill, F. B. (1997). Life Before Model Systems: General Zoology at August Weismann's. *Variety 268*(December 1995), 260–268.

Civenni, G., A. Walter, N. Kobert, D. Mihic-Probst, M. Zipser, B. Belloni, B. Seifert, H. Moch, R. Dummer, M. van den Broek, and L. Sommer (2011). Human CD271-positive melanoma stem cells associated with metastasis establish tumor heterogeneity and long-term growth. *Cancer research 71*(8), 3098–109.

Clarke, M. F., J. E. Dick, P. B. Dirks, C. J. Eaves, C. H. M. Jamieson, D. L. Jones, J. Visvader, I. L. Weissman, and G. M. Wahl (2006). Cancer stem cells–perspectives on current status and future directions: AACR Workshop on cancer stem cells. *Cancer research 66*(19), 9339–44.

Colman, A. (2008). Induced pluripotent stem cells and human disease. *Cell stem cell 3*(3), 236–7.

Colman, A. and O. Dreesen (2009). Pluripotent stem cells and disease modeling. *Cell stem cell 5*(3), 244–7.

Committee on a Framework for Development a New Taxonomy of Disease (2011). *Toward Precision Medicine : Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press.

Committee on Models for Biomedical Research (1985). Models for Biomedical Research: A New Perspective. Technical report, National Research Council (NRC).

Committee on the Use of Laboratory Animals (1988). Use of laboratory animals in biomedical and behavioral research. Technical report, National Research Council.

Conti, L. and E. Cattaneo (2010). Neural stem cell systems: physiological players or in vitro entities? *Nature reviews. Neuroscience 11*(3), 176–87.

Corbett, T., L. Polin, P. LoRusso, F. Valeriote, C. Panchapor, S. Pugh, K. White, J. Knight, L. Demchik, J. Jones, L. Jones, and L. Lisow (2004). In Vivo Methods for Screening and Preclinical Testing: Use of Rodent Solid Tumors for Drug Discovery. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 99–124. Humana Press.

Coste, J., B. Cochand-Priollet, and P. D. Cremoux (2003). Cross sectional study of conventional cervical smear, monolayer cytology, and human papillomavirus DNA testing for cervical cancer screening. *British Medical Journal 326*(April), 1–5.

Creager, A. N. H., E. Lunbeck, and M. N. Wise (2007). *Science without Laws: Model Systems, Cases, Exemplary Narratives*. Duke University Press.

Creighton, C., R. Kuick, D. E. Misek, D. S. Rickman, F. M. Brichory, J.-M. Rouillard, G. S. Omenn, and S. Hanash (2003). Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice. *Genome biology 4*(7), R46.

Cronbach, L. J. and P. E. Meehl (1955). Construct validity in psychological tests. *Psychological Bulletin 52*, 281–302.

Cuthbert, B. N. and T. R. Insel (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine 11*(1), 126.

da Costa, N. C. and S. French (2003). *Science and partial truth*. Oxford University Press.

Della Negra, M., P. Jenni, and T. S. Virdee (2012). Journey in the search for the Higgs boson: the ATLAS and CMS experiments at the Large Hadron Collider. *Science 338*(6114), 1560–8.

DeRose, Y. S., G. Wang, Y.-C. Lin, P. S. Bernard, S. S. Buys, M. T. W. Ebbert, R. Factor, C. Matsen, B. a. Milash, E. Nelson, L. Neumayer, R. L. Randall, I. J. Stijleman, B. E. Welm, and A. L. Welm (2011). Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nature medicine 17*(11), 1514–20.

Derrida, J. (1967). *De la Grammatologie*. Les Éditions de Minuit.

DeVita, V. T. and E. Chu (2008). A history of cancer chemotherapy. *Cancer research 68*(21), 8643–53.

Dick, J. E. (2003). Breast cancer stem cells revealed. *Proceedings of the National Academy of Sciences of the United States of America 100*(7), 3547–3549.

Ding, Q., Y.-K. Lee, E. Schaefer, D. Peters, A. Veres, K. Kim, N. Kuperwasser, D. Motola, T. Meissner, W. Hendriks, M. Trevisan, R. Gupta, A. Moisan, E. Banks, M. Friesen, R. Schinzel, F. Xia, A. Tang, Y. Xia, E. Figueroa, A. Wann, T. Ahfeldt, L. Daheron, F. Zhang, L. Rubin, L. Peng, R. Chung, K. Musunuru, and C. Cowan (2012). A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Cell Stem Cell*, 1–14.

Drouin-Ouellet, J. and R. A. Barker (2012). Parkinson's Disease in a Dish: What Patient Specific-Reprogrammed Somatic Cells Can Tell Us about Parkinson's Disease, If Anything? *Stem Cells International 2012*, 1–10.

Duhem, P. M. M. (1906). *La théorie physique: son objet, et sa structure*. Chevalier et Rivière.

Eagle, H. (1958). Discussion of Part II. *Annals of the New York Academy of Sciences 76*, 542–555.

Ehret, G. B., P. B. Munroe, K. M. Rice, M. Bochud, et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature 478*(7367), 103–9.

European Comission (2010). Sixth Report on the Statistics on the Number of Animals used for Experimental and other Scientific Purposes in the Member States of the European Union SEC(2010) 1107. Technical report.

Ewing, J. (1919). *Neoplastic diseases, a text-book on tumors*. W.B. Saunders Company.

Farah, I. O., M. Ngotho, T. Kariuki, M. Jeneby, L. Irura, N. Maina, J. Kagira, M. Gicheru, and J. Hau (2002). Animal Models for Tropical Parasitic Diseases. In J. Hau (Ed.), *Handbook of Laboratory Animal Science Volume III: Animal Models*, pp. 169–224. CRC Press.

Ferrero, G. B., C. Howald, L. Micale, E. Biamino, B. Augello, C. Fusco, M. G. Turturo, S. Forzano, A. Reymond, and G. Merla (2010). An atypical 7q11.23 deletion in a normal IQ Williams-Beuren syndrome patient. *European journal of human genetics : EJHG 18*(1), 33–8.

Freitas, B. C. G., C. a. Trujillo, C. Carromeu, M. Yusupova, R. Herai, and A. R. Muotri (2012). Stem cells and modeling of autism spectrum disorders. *Experimental neurology*.

Frigg, R. and S. Hartmann (2012). Models in science. *Standfor Encyclopedia of Philosophy*.

Fuller, J. L. and W. R. Thompson (1960). *Behavior genetics*. Wiley.

Funk, C. (1915). The transplantation of tumors to foreign species. *The Journal of Experimental Medicine XXI*.

Gachelin, G. (Ed.) (2006). *Les organismes modèles dans la recherche médicale*. Presses Universitaires de France.

Gad, S. C. (2007). *Animal Models in Toxicology* (2nd ed.). CRC Press.

Gayon, J. (2006). Les organismes modèles en biologie et en médecine. In G. Gachelin (Ed.), *Les organismes modèles dans la recherche médicale*. Presses Universitaires de France.

Geertz, C. (2000). *The interpretation of cultures*. Basic Books.

Gellhorn, A. and E. Hirschberg (1955). Investigations of diverse systems for cancer chemotherapy screening. *Cancer research 3*(1).

Gerlinger, M., A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, and C. Swanton (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine 366*(10), 883–892.

Germain, P.-L. (2012). Cancer cells and adaptive explanations. *Biology & Philosophy 27*(6), 785–810.

Germain, P.-L. (forthcoming a). From Replica to Instruments: Animal Models in Contemporary Biomedical Research. *History and Philosophy of the Life Sciences*.

Germain, P.-L. (forthcoming b). Living instruments and theoretical terms. In M. C. Galavotti, S. Hartmann, M. Weber, W. Gonzalez, D. Dieks, and T. Uebel (Eds.), *New Directions in the Philosophy of Science*. Springer.

Giere, R. (2004). How models are used to represent reality. *Philosophy of Science 71*(December), 742–752.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology & Philosophy 21*(5), 725–740.

Goldstein, L. S. B. (2012). New frontiers in human cell biology and medicine: Can pluripotent stem cells deliver? *The Journal of cell biology 199*(4), 577–81.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.

Goodman, N. (1976 [1968]). *Languages of Art: An Approach to a Theory of Symbols* (2nd ed.). Hackett Pub Co.

Gradmann, C. (2006). Maladies expérimentales. Les Expériences sur l'animal aux débuts de la bactériologie médicale. In G. Gachelin (Ed.), *Les organismes modèles dans la recherche médicale*. Presses Universitaires de France.

Graf, T. (2011). Historical Origins of Transdifferentiation and Reprogramming. *Cell Stem Cell 9*(6), 504–516.

Graham, G. (2010). Behaviorism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2010 ed.).

Greek, R. and N. Shanks (2011). Complex systems, evolution, and animal models. *Studies in history and philosophy of biological and biomedical sciences 42*(4), 542–4.

Greene, H. S. (1951). A conception of tumor autonomy based on transplantation studies: a review. *Cancer research 11*(12), 899–903.

Greene, H. S. (1952). The significance of the heterologous transplantability of human cancer. *Cancer*, 24–44.

Greene, H. S. N. (1948). Identification of malignant tissues. *JAMA : the journal of the American Medical Association 137*(16), 1364–1366.

Griesemer, J. (1990). Material models in biology. *PSA: Proceedings of the Biennial meeting of the Philosophy of Science Association 2*, 79–93.

Griesemer, J. R. (1992). The Role of Instruments in the Generative Analysis of Science. In *The right tools for the job: at work in the Twentieth century life sciences*, pp. 47–76.

Grskovic, M., A. Javaherian, B. Strulovici, and G. Q. Daley (2011). Induced pluripotent stem cells - opportunities for disease modelling and drug discovery. *Nature reviews. Drug discovery* (November).

Gupta, V. and K. D. Poss (2012). Clonally dominant cardiomyocytes direct heart morphogenesis. *Nature 484*(7395), 479–484.

Gurdon, J. B. and J. a. Byrne (2003). The first half-century of nuclear transplantation. *Bioscience reports 24*(4-5), 545–57.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Hacking, I. (1992). The Self-Vindication of the Laboratory Sciences. In A. Pickering (Ed.), *Science as practice and culture*, pp. 29–64. University of Chicago Press.

Han, S. S. W., L. a. Williams, and K. C. Eggan (2011). Constructing and deconstructing stem cell models of neurological disease. *Neuron 70*(4), 626–644.

Hanna, J., M. Wernig, S. Markoulaki, C.-W. Sun, A. Meissner, J. P. Cassady, C. Beard, T. Brambrink, L.-C. Wu, T. M. Townes, and R. Jaenisch (2007). Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science 318*(5858), 1920–3.

Hauskeller, C. and S. Weber (2011). Framing pluripotency: iPS cells and the shaping of stem cell science. *New Genetics and Society* (May 2012), 37–41.

Hayflick, L. (1965). The limited in vitro lifetime of human diploid cell strains. *Experimental Cell Research 37*(3), 614–636.

Hayflick, L. (2000). The illusion of cell immortality. *British journal of cancer 83*(7), 841–6.

Heidelberger, M. (2003). Theory-Ladenness and Scientific Instruments in Experimentation. In H. Radder (Ed.), *The Philosophy of Scientific Experimentation*, pp. 138–151. University of Pittsburgh Press.

Hekzog, M. (1902). On tumor transplantation and inoculation. *The Journal of medical research 1*(6), 74–84.

Hempel, C. G. (1965). *Apects of Scientific Explanation: And Other Essays in the Philosophy of Science*. The Free Press.

Hinson, J. T., K. Nakamura, and S. M. Wu (2012). Induced pluripotent stem cell modeling of complex genetic diseases. *Drug Discovery Today: Disease Models xxx*(xx), 2–7.

Hope, K. J., L. Jin, and J. E. Dick (2004). Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nature immunology 5*(7), 738–43.

Hubbard, E. J. A. (2007). Model Organisms as Powerful Tools for Biomedical Research. In A. N. H. Creager, E. Lunbeck, and M. N. Wise (Eds.), *Science without Laws: Model Systems, Cases, Exemplary Narratives*, pp. 59–72. Duke University Press.

Huber, L. and L. Keuck (2013). Mutant mice: Experimental organisms as materialised models in biomedicine. *Studies in History and Philosophy of Biological and Biomedical Sciences 44*(3), 385–391.

ILAR and NRC (1969). Animal Models for Biomedical Research II. Technical report, National Academy of Sciences.

ILAR and NRC (1998). Biomedical Models and Resources: Current Needs and Future Opportunities. Technical report, Committee on New and Emerging Models in Biomedical and Behavioral Research, Institute for Laboratory Animal Research, Commission on Life Sciences, National Research Council.

Institute of Laboratory Animal Resources (ILAR) (1977). The Future of Animals, Cells, Models and Systems in Research, Development, Education and Testing. National Research Council (NRC): National Academy of Sciences.

International Committee on Laboratory Animals (1971). Defining the Laboratory Animal. In *IV Symposium (1969)*. International Committe on Laboratory Animals & Institute of Laboratory Animal Resources, National Research Council: National Academy of Sciences.

Israel, M. a., S. H. Yuan, C. Bardy, S. M. Reyna, Y. Mu, C. Herrera, M. P. Hefferan, S. Van Gorp, K. L. Nazor, F. S. Boscolo, C. T. Carson, L. C. Laurent, M. Marsala, F. H. Gage, A. M. Remes, E. H. Koo, and L. S. B. Goldstein (2012). Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature*.

Jablonka, E. (1996). Do cells show off? Somatic selection and the nature of intercellular signalling. *Tree 11*(10), 395–396.

Jucker, M. (2010). The benefits and limitations of animal models for translational research in neurodegenerative diseases. *Nature medicine 16*(11), 1210–4.

Keating, P. and A. Cambrosio (2012). *Cancer on Trial: Oncology as a New Style of Practice*. University of Chicago Press.

Keller, E. F. (2000). Models Of and Models For : Theory and Practice in Contemporary Biology. *Philosophy of Science 67*, S72–S86.

Keller, E. F. (2002). *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines*. Harvard University Press.

Kim, J., J. Hoffman, and R. Alpaugh (2013). An iPSC Line from Human Pancreatic Ductal Adenocarcinoma Undergoes Early to Invasive Stages of Pancreatic Cancer Progression. *Cell Reports*, 1–12.

Kim, T.-K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. a. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature 465*(7295), 182–7.

Kincaid, H. (1990). Molecular Biology and the Unity of Science. *Philosophy of Science 57*(4), 575–593.

Kirk, R. G. W. (2012). "Standardization through Mechanization": Germ-Free Life and the Engineering of the Ideal Laboratory Animal. *Technology and Culture 53*(1), 61–93.

Kitcher, P. (2001). Real realism: the Galilean strategy. *Philosophical Review 110*(2), 151–197.

Kitcher, P. (2003). *Science, Truth, and Democracy*. Oxford University Press, USA.

Knight, A. (2008). Systematic reviews of animal experiments demonstrate poor contributions toward human healthcare. *Reviews on recent clinical trials 3*(2), 89–96.

Knight, A. (2011). *The costs and benefits of animal experiments*. Palgrave Macmillan.

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science Part A 42*(2), 262–271.

Kohler, R. E. (1991). Systems of Production: Drosophila, Neurospora, and Biochemical Genetics. *Historical Studies in the Physical and Biological Sciences 22*(1), 87– 130.

Ladewig, J., P. Koch, and O. Brüstle (2013). Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nature reviews. Molecular cell biology 14*(4), 225–36.

LaFollette, H. and N. Shanks (1994). Animal experimentation: The legacy of Claude Bernard. *International Studies in the Philosophy of Science 8*(3), 195–210.

LaFollette, H. and N. Shanks (1995). Two models of models in biomedical research. *The Philosophical Quarterly 45*(179).

LaFollette, H. and N. Shanks (1996). *Brute science. Dilemmas of animal experimentation.* Routledge.

Lancaster, M. a., M. Renner, C.-A. Martin, D. Wenzel, L. S. Bicknell, M. E. Hurles, T. Homfray, J. M. Penninger, A. P. Jackson, and J. a. Knoblich (2013). Cerebral organoids model human brain development and microcephaly. *Nature*.

Landecker, H. (2007). *Culturing life: How cells became technologies*. Harvard University Press.

Lasagna, L. (1958). Extrapolation to studies in man. *Annals of the New York Academy of Sciences*.

Ledford, H. (2008). The full cycle. *Nature 453*, 843–845.

Lee, G. and L. Studer (2010). Induced pluripotent stem cell technology for the study of human disease. *Nature methods 7*(1), 25–7.

Lenski, R. E. and M. Travisano (1994). Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America 91*(15), 6808–14.

Leonelli, S. (2007). Performing abstraction: two ways of modelling Arabidopsis thaliana. *Biology & Philosophy 23*(4), 509–528.

Levins, R. (1966). The strategy of model building in population biology. *American scientist 54*(4).

Loeb, L. (1945). *The Biological Basis of Individuality*. Charles C. Thomas Publisher.

Logan, C. A. (2002). Before There Were Standards: The Role of Test Animals in the Production of Empirical Generality in Physiology. *Journal of the History of Biology*, 329–363.

Mäki, U. (2005). Models are experiments, experiments are models. *Journal of Economic Methodology 12*(2), 303–315.

Mandler, G. (2002). Origins of the cognitive (r) evolution. *Journal of the History of the Behavioral Sciences 38*(4), 339–353.

Marchetto, M. and F. Gage (2012). Modeling Brain Disease in a Dish: Really? *Cell Stem Cell 10*(6), 642–645.

Marchetto, M. C. N., A. R. Muotri, Y. Mu, A. M. Smith, G. G. Cezar, and F. H. Gage (2008). Non-cell-autonomous effect of human SOD1 G37R astrocytes on motor neurons derived from human embryonic stem cells. *Cell stem cell 3*(6), 649–57.

Mariani, J., M. V. Simonini, D. Palejev, L. Tomasini, G. Coppola, a. M. Szekely, T. L. Horvath, and F. M. Vaccarino (2012). Modeling human cortical development in vitro using induced pluripotent stem cells. *Proceedings of the National Academy of Sciences*, 1–7.

Marincola, F. M. (2003). Translational Medicine: A two-way road. *Journal of translational medicine 1*(1), 1.

Mattis, V. B. and C. N. Svendsen (2011). Induced pluripotent stem cells: a new revolution for clinical neurology? *Lancet neurology 10*(4), 383–94.

Maugeri, P. and A. Blasimme (2011). Humanised models of cancer in molecular medicine: the experimental control of disanalogy. *History and philosophy of the life sciences 33*, 603–622.

Mauro, R. and M. Kubovy (1992). Caricature and face recognition. *Mem. Cogn. 20*, 433–440.

Mayet, M. (1902). Production du Cancer chez les Rats blancs par Introduction dans leurs economies des Substances constituantes des Tumeurs malignes de l'Homme. *Gazette hebdomadaire de médecine et de chirurgie 6*.

Merkle, F. and K. Eggan (2013). Modeling Human Disease with Pluripotent Stem Cells: from Genome Association to Function. *Cell Stem Cell 12*(6), 656–668.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist 50*(9), 741–749.

Meunier, R. (2011). *Thick and thin characters: Organismal form and representational practice in embryology and genetics.* Ph. D. thesis, European School of Molecular Medicine (SEMM) and University of Milan.

Meunier, R. (2012). Stages in the development of a model organism as a platform for mechanistic models in developmental biology: Zebrafish, 1970–2000. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43*(2), 522–531.

Millan, M. J. (2008). The Discovery and Development of Pharmacotherapy for Psychiatric Disorders: A Critical Survey of Animal and Translational Models and Perspectives for Their Improvement. In R. McArthur and F. Borsini (Eds.), *Animal and translational models for CNS drug discovery. Volume 1: Psychiatric Disorders*, pp. 1–57. Academic Press.

Morange, M. (1997). From the Regulatory Vision of Cancer to the Oncogene Paradigm, 1975-1985. *Journal of the History of Biology 30*, 1–29.

Morgan, M. S. (2003). Experiments without Material Intervention: Model experiments, virtual experiments, and virtually experiments. In *The Philosophy of Scientific Experimentation*, pp. 216–235. University of Pittsburgh Press.

Morgan, M. S. and M. Morrison (Eds.) (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.

Nelson, N. C. (2012). Modeling mouse, human, and discipline: Epistemic scaffolds in animal behavior genetics. *Social Studies of Science*.

Newton-Cheh, C., T. Johnson, V. Gateva, M. D. Tobin, et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nature genetics 41*(6), 666–76.

Nguyen, H. N., B. Byers, B. Cord, A. Shcheglovitov, J. Byrne, P. Gujar, K. Kee, B. Schu, R. E. Dolmetsch, W. Langston, T. D. Palmer, and R. R. Pera (2011). LRRK2 Mutant iPSC-Derived DA Neurons Demonstrate Increased Susceptibility to Oxidative Stress. *Cell Stem Cell* (8), 267–280.

Nishikawa, S.-i., R. A. Goldstein, and C. R. Nierras (2008). The promise of human induced pluripotent stem cells for research and therapy. *Nature Reviews Molecular Cell Biology 9*(September), 725–729.

Nowotny, H. and G. Testa (2011). *Naked Genes: Reinventing the Human in the Molecular Age*. The MIT Press.

Olson, H., G. Betton, D. Robinson, K. Thomas, A. Monro, G. Kolaja, P. Lilly, J. Sanders, G. Sipes, W. Bracken, M. Dorato, K. Van Deun, P. Smith, B. Berger, and A. Heller (2000). Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory toxicology and pharmacology RTP 32*(1), 56–67.

Orzack, S. H. and E. Sober (1994). Optimality models and the test of adaptationism. *American Naturalist 143*(3), 361–380.

Pagé, M. (2004). High-Volume Screening. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 3–22. Humana Press.

Pasca, S. P., T. Portmann, I. Voineagu, M. Yazawa, A. Shcheglovitov, A. M. Pasca, B. Cord, T. D. Palmer, S. Chikahisa, S. Nishino, J. A. Bernstein, J. Hallmayer, D. H. Geschwind, R. E. Dolmetsch, S. P. Paşca, and A. M. Paşca (2011). Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. *Nature Medicine 17*(12), 1657–1663.

Pickering, S. J., S. L. Minger, M. Patel, H. Taylor, C. Black, C. J. Burns, A. Ekonomou, and P. R. Braude (2005). Generation of a human embryonic stem cell line encoding the cystic fibrosis mutation ΔF508, using preimplantation genetic diagnosis. *Reproductive BioMedicine Online 10*(3), 390–397.

Piotrowska, M. (2012). From humanized mice to human disease: guiding extrapolation from model to target. *Biology & Philosophy*.

Porsolt, R. D., A. Bertin, and M. Jalfre (1977). Behavioral despair in mice: a primary screening test for antidepressants. *Archives Internationales De Pharmacodynamie Et De Therapie 229*(2), 327–336.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *Philosophical Review 60*(1), 20–43.

Quintana, E., M. Shackleton, M. S. Sabel, D. R. Fullen, T. M. Johnson, and S. J. Morrison (2008). Efficient tumour formation by single human melanoma cells. *Nature 456*(7222), 593–598.

Radder, H. (Ed.) (2003). *The Philosophy of Scientific Experiment*. University of Pittsburgh Press.

Rader, K. (2004). *Making Mice: Standardizing Animals for American Biomedical Research, 1900-1955*. Princeton University Press.

Rais, Y., A. Zviran, S. Geula, O. Gafni, E. Chomsky, S. Viukov, A. A. Mansour, I. Caspi, V. Krupalnik, M. Zerbib, I. Maza, N. Mor, D. Baran, L. Weinberger, D. a. Jaitin, D. Lara-Astiaso, R. Blecher-Gonen, Z. Shipony, Z. Mukamel, T. Hagai, S. Gilad, D. Amann-Zalcenstein, A. Tanay, I. Amit, N. Novershtern, and J. H. Hanna (2013). Deterministic direct reprogramming of somatic cells to pluripotency. *Nature 502*(7469), 65–70.

Ratcliff, W. C., R. F. Denison, M. Borrello, and M. Travisano (2012). Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences of the United States of America 109*(5), 1595–600.

Rheinberger, H.-J. (1995a). From Experimental Systems to Cultures of Experimentation. In G. Wolters and J. G. Lennox (Eds.), *Concepts, Theories, and Rationality in the Biological Sciences*, pp. 107–122. University of Pittsburgh Press.

Rheinberger, H. J. (1995b). From microsomes to ribosomes: "strategies" of "representation". *Journal of the history of biology 28*(1), 49–89.

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press.

Rheinberger, H.-J. (2000). Beyond nature and culture: modes of reasoning in the age of molecular biology and medicine. In M. Lock, A. Young, and A. Cambrosio (Eds.), *Living and working with the new medical technologies: intersections of inquiry*, pp. 19–30. Cambridge University Press.

Rheinberger, H.-J. (2006). Réflexions sur les organismes modèles dans la recherche biologique au XXe siècle. In G. Gachelin (Ed.), *Les organismes modèles dans la recherche médicale*. Presses Universitaires de France.

Rheinberger, H.-J. (2010). *An Epistemology of the Concrete: Twentieth-Century Histories of Life*. Duke University Press.

Rheinberger, H.-J. (2011a). Consistency from the perspective of an experimental systems approach to the sciences and their epistemic objects. *Manuscrito 34*(1), 307–321.

Rheinberger, H.-j. (2011b). Recent Orientations and Reorientations in the Life Sciences. In M. Carrier and A. Nordmann (Eds.), *Science in the Context of Application*, Volume 274, pp. 161–168. Springer.

Rhodes, G., S. Brennan, and S. Carey (1987). Identification and ratings of caricatures: implications for mental representations of faces. *Cogn. Psychol.* (19), 473–497.

Roelcke, V. (2009). Tiermodell und Menschenbild. Konfigurationen der epistemologischen und ethischen Mensch-Tier-Grenzziehung in der Humanmedizin zwischen 1880 und 1945. In B. Griesecke, M. Krause, N. Pethes, and K. Sabisch (Eds.), *Kulturgeschichte des Menschenversuchs im 20.Jahrhundert*. Suhrkamp.

Rossant, J. (2007). The magic brew. *Nature News 448*(July), 2–4.

Saha, K. and J. B. Hurlbut (2011). Disease modeling using pluripotent stem cells: making sense of disease from bench to bedside. *Swiss medical weekly 141*(February), w13144.

Salsburg, D. (1983). The lifetime feeding study in mice and rats–an examination of its validity as a bioassay for human carcinogens. *Fundamental and applied toxicology : official journal of the Society of Toxicology 3*(1), 63–7.

Santoriello, C., E. Gennaro, V. Anelli, M. Distel, A. Kelly, R. W. Köster, A. Hurlstone, and M. Mione (2010). Kita driven expression of oncogenic HRAS leads to early onset and highly penetrant melanoma in zebrafish. *PloS one 5*(12), e15170.

Sasai, Y., M. Ogushi, T. Nagase, and S. Ando (2008). Bridging the gap from frog research to human therapy: a tale of neural differentiation in Xenopus animal caps and human pluripotent cells. *Development, growth & differentiation 50 Suppl 1*, S47–55.

Scannell, J. W., A. Blanckley, H. Boldon, and B. Warrington (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews. Drug discovery 11*(3), 191–200.

Schaffner, K. F. (1993). *Discovery and Explanation in Biology and Medicine*. University of Chicago Press.

Schatton, T., G. F. Murphy, N. Y. Frank, K. Yamaura, A. M. Waaga-Gasser, M. Gasser, Q. Zhan, S. Jordan, L. M. Duncan, C. Weishaupt, R. C. Fuhlbrigge, T. S. Kupper, M. H. Sayegh, and M. H. Frank (2008). Identification of cells initiating human melanomas. *Nature 451*(7176), 345–349.

Schatton, T., U. Schütte, N. Y. Frank, Q. Zhan, A. Hoerning, S. C. Robles, J. Zhou, F. S. Hodi, G. C. Spagnoli, G. F. Murphy, and M. H. Frank (2010). Modulation of T-cell activation by malignant melanoma initiating cells. *Cancer Research 70*(2), 697–708.

Schepartz, S. A., B. J. Abbott, and J. Leiter (1967). Screening Data from the Cancer Chemotherapy National Service Center Screening Laboraties. XLIII. *Cancer research 27*, 737–911.

Scholl, R. and T. Räz (2012). Modeling causal structures. *European Journal for Philosophy of Science 3*(1), 115–132.

Shanks, N. and C. R. Greek (2009). *Animal models in light of evolution*. BrownWalker Press.

Shanks, N., R. Greek, and J. Greek (2009). Are animal models predictive for humans? *Philosophy, ethics, and humanities in medicine 4*, 2.

Shelley, C. (2010). Why test animals to treat humans? On the validity of animal models. *Studies in history and philosophy of biological and biomedical sciences 41*(3), 292–299.

Shi, Y., P. Kirwan, J. Smith, H. P. C. Robinson, and F. J. Livesey (2012). Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nature neuroscience* (February).

Shibata, D. (2012). Heterogeneity and Tumor History. *Science 336*(304).

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer 6*(10), 813–23.

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals*. Number 0. Random House.

Singer, T., M. J. McConnell, M. C. N. Marchetto, N. G. Coufal, and F. H. Gage (2010). LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in Neurosciences 33*(8), 345–354.

Singh, S. K., I. D. Clarke, M. Terasaki, V. E. Bonn, C. Hawkins, J. Squire, and P. B. Dirks (2003). Identification of a cancer stem cell in human brain tumors. *Cancer Research 63*(18), 5821–8.

Snell, G. D. (1964). The terminology of tissue transplantation. *Transplantation 2*, 655.

Sonnenschein, C. and A. M. Soto (2008). Theories of carcinogenesis: an emerging perspective. *Seminars in cancer biology 18*(5), 372–377.

Spear, B. B., M. Heath-Chiozzi, and J. Huff (2001). Clinical application of pharmacogenetics. *Trends in Molecular Medicine 7*(5), 201–204.

Steel, D. (2010). A new approach to argument by analogy: extrapolation and chain graphs. *Philosophy of Science 77*(5), 1058–1069.

Steel, D. P. (2008). *Across The Boundaries: Extrapolation in Biology and Social Science*. Cambridge University Press.

Suarez, M. (2004). An Inferential Conception of Scientific Representation. *Philosophy of Science 71*(5), 767–779.

Suárez, M. (2010). Scientific Representation. *Philosophy Compass 5*(1), 91–101.

Suppe, F. (1974). *Structure of Scientific Theories*. University of Illinois Press.

Suppes, P. (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese* (12), 287–301.

Suvà, M. L., N. Riggi, and B. E. Bernstein (2013). Epigenetic reprogramming in cancer. *Science (New York, N.Y.) 339*(6127), 1567–70.

Swoyer, C. (1991). Structural Representation and Surrogative Reasoning. *Synthese* (87), 449–508.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell 131*(5), 861–72.

Takahashi, K. and S. Yamanaka (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell 126*(4), 663–676.

Testa, G. (2009). What to do with the Grail now that we have it? iPSCs, potentiality, and public policy. *Cell stem cell 5*(4), 358–9.

Testa, G. (forthcoming). Democracies of Stemness: Stem Cell Technologies from Generation to Regeneration. In F. Calegari and C. Waskow (Eds.), *Stem Cells: From Basic Research to Therapy, Volume Two: Tissue Homeostasis and Regeneration during Adulthood, Applications, Legislation and Ethics*, Chapter 13, pp. 401–427. CRC Press.

Thambi, P. and E. A. Sausville (2004). Working With the National Cancer Institute. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 339–350. Humana Press.

The Nobel Committee for Physiology or Medicine (2012). The 2012 Nobel Prize in Physiology or Medicine. Technical report.

Theil, F.-P., T. W. Guentert, S. Haddad, and P. Poulin (2003). Utility of physiologically based pharmacokinetic models to drug development and rational drug discovery candidate selection. *Toxicology letters 138*(1-2), 29–49.

Tian, X., J. Azpurua, C. Hine, A. Vaidya, M. Myakishev-Rempel, J. Ablaeva, Z. Mao, E. Nevo, V. Gorbunova, and A. Seluanov (2013). High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature*, 1–6.

Topczewska, J. M., L.-M. Postovit, N. V. Margaryan, A. Sam, A. R. Hess, W. W. Wheaton, B. J. Nickoloff, J. Topczewski, and M. J. C. Hendrix (2006). Embryonic and tumorigenic pathways converge via Nodal signaling: role in melanoma aggressiveness. *Nature medicine 12*(8), 925–32.

Towbin, A. (1951). The heterologous transplantation of human tumors. *Cancer Research 11*, 716–722.

Unternaehrer, J. J. and G. Q. Daley (2011). Induced pluripotent stem cells for modelling human diseases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences 366*(1575), 2274–85.

Urbach, A., O. Bar-Nur, G. Q. Daley, and N. Benvenisty (2010). Differential modeling of fragile X syndrome by human embryonic stem cells and induced pluripotent stem cells. *Cell Stem Cell* (6), 407–411.

Valent, P., C. Eaves, D. Bonnet, R. De Maria, T. Lapidot, M. Copland, J. V. Melo, C. Chomienne, F. Ishikawa, J. J. Schuringa, G. Stassi, B. Huntly, H. Herrmann, J. Soulier, A. Roesch, G. J. Schuurhuis, S. Wöhrer, M. Arock, J. Zuber, S. Cerny-Reiterer, H. E. Johnsen, and M. Andreeff (2012). Cancer stem cell definitions and terminology: the devil is in the details. *Nature Reviews Cancer 12*(11), 767–775.

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Clarendon Press.

Varmus, H. E. (2010). NCI town hall meeting, July 12, 2010, Natcher Center. Accessible at http://videocast.nih.gov/launch.asp?16014.

Vignais, P. and P. Vignais (2010). *Discovering Life, Manufacturing Life: How the experimental method shaped life sciences*. Springer.

Visvader, J. and G. Lindeman (2012). Cancer Stem Cells: Current Status and Evolving Complexities. *Cell Stem Cell 10*(6), 717–728.

Vogel, G. and D. Normile (2012). Reprogrammed Cells Earn Biologists Top Honor. *Science News 338*(October), 178–179.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature 118*(2972), 558–560.

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *Journal du Conseil - Conseil International pour l'Exploration de la Mer 3*(1), 3–51.

Wahlsten, D. (2012). The hunt for gene effects pertinent to behavioral traits and psychiatric disorders: From mouse to human. *Developmental Psychobiology*, n/a–n/a.

Warren, L., P. D. Manos, T. Ahfeldt, Y.-H. Loh, H. Li, F. Lau, W. Ebina, P. K. Mandal, Z. D. Smith, A. Meissner, G. Q. Daley, A. S. Brack, J. J. Collins, C. Cowan, T. M. Schlaeger, and D. J. Rossi (2010). Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell stem cell 7*(5), 618–30.

Waters, C. K. (2012). Experimental modeling as a form of theoretical modeling. *Paper presented at PSA 2012 (Biennial Meeting of the Philosophy of Science Association), San Diego, 15-17 November 2012*.

Waud, W. R. (2004). Murine L1210 and P388 Leukemias. In B. A. Teicher and P. A. Andrews (Eds.), *Anticancer Drug Development Guide. Preclinical Screening, Clinical Trials, and Approval* (2 ed.)., pp. 79–98. Humana Press.

Weber, M. (2005). *The Philosophy of Experimental Biology*. Cambridge University Press.

Weber, M. (2011). Experimentation versus Theory Choice : A Social-Epistemological Approach. In H. B. Schmid, D. Sirtes, and M. Weber (Eds.), *Collective Epistemology*, pp. 203–225. Ontos Verlag.

Weber, M. (2012). Experimental Modeling: Exemplification and Representation as Theorizing Strategies. *Paper presented at PSA 2012 (Biennial Meeting of the Philosophy of Science Association), San Diego, 15-17 November 2012*.

Weisberg, M. (2007). Who is a Modeler? *The British Journal for the Philosophy of Science 58*(2), 207–233.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Wernig, M., A. Meissner, R. Foreman, T. Brambrink, M. Ku, K. Hochedlinger, B. E. Bernstein, and R. Jaenisch (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature 448*(7151), 318–24.

White, J. K., A.-K. Gerdin, N. a. Karp, E. Ryder, M. Buljan, J. N. Bussell, J. Salisbury, S. Clare, N. J. Ingham, C. Podrini, R. Houghton, J. Estabel, J. R. Bottomley, D. G. Melvin, D. Sunter, N. C. Adams, D. Tannahill, D. W. Logan, D. G. Macarthur, J. Flint, V. B. Mahajan, S. H. Tsang, I. Smyth, F. M. Watt, W. C. Skarnes, G. Dougan, D. J. Adams, R. Ramirez-Solis, A. Bradley, and K. P. Steel (2013). Genome-wide Generation

and Systematic Phenotyping of Knockout Mice Reveals New Roles for Many Genes. *Cell 154*(2), 452–64.

White, R., K. Rose, and L. Zon (2013). Zebrafish cancer: the state of the art and the path forward. *Nature Reviews Cancer 13*(9), 624–636.

Willner, P. (1984). The validity of animal models of depression. *Psychopharmacology Berl 83*(1), 1–16.

Willner, P. (1991). Methods for assessing the validity of animal models of human psychopathology. *Animal models in psychiatry, I 18*, 1–23.

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Witkowski, J. A. (1983). Experimental pathology and the origins of tissue culture: Leo Loeb's contribution. *Medical History 27*(03), 269–288.

Woglom, W. H. (1913). *Studies in cancer and allied subjects. The study of experimental cancer.* New York: Columbia University Press.

Wu, G., N. Liu, I. Rittelmeyer, A. D. Sharma, M. Sgodda, H. Zaehres, M. Bleidiß el, B. Greber, L. Gentile, D. W. Han, C. Rudolph, D. Steinemann, A. Schambach, M. Ott, H. R. Schöler, and T. Cantz (2011). Generation of Healthy Mice from Gene-Corrected Disease-Specific Induced Pluripotent Stem Cells. *PLoS biology 9*(7), e1001099.

Xu, M., Z. Sulkowski, P. Parekh, and A. Khan (2013). Effects of Perinatal Lipopolysaccharide (LPS) Exposure on the Developing Rat Brain; Modeling the Effect of Maternal Infection on the Developing Human CNS. *The Cerebellum 12*(4), 572–586.

Yamanaka, S. (2012). Induced Pluripotent Stem Cells: Past, Present, and Future. *Cell Stem Cell 10*(6), 678–684.

Young, J. E. and L. S. B. Goldstein (2012). Alzheimer's Disease in a Dish: Promises and Challenges of Human Stem Cell Models. *Human Molecular Genetics* (858), 1–21.

Zhao, X.-y., W. Li, Z. Lv, L. Liu, M. Tong, T. Hai, J. Hao, C.-l. Guo, Q.-w. Ma, L. Wang, F. Zeng, and Q. Zhou (2009). iPS cells produce viable mice through tetraploid complementation. *Nature 461*(7260), 86–90.

Zondek, B. (1928). Die Schwangerschaftsdiagnose aus dem Harn durch Nachweis des Hypophysenvorderlappenhormons. *Die Naturwissenschaften 51*, 1088–1090.